Universidade FUMEC
Faculdade de Ciências Empresariais
Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento

Anotação Semântica Automática do Currículo Lattes Utilizando Linked Open Data

Walison Dias da Silva

Belo Horizonte
2016

Walison Dias da Silva

Anotação Semântica Automática do Currículo Lattes Utilizando Linked Open Data

Dissertação de mestrado apresentado ao Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento como parte dos requisitos para a obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento. Área de concentração: Gestão de Sistemas de Informação e do Conhecimento. Linha de pesquisa: Tecnologia e Sistemas de Informação.

Orientador: Prof. Dr. Fernando Silva Parreiras

Belo Horizonte

2016

Agradecimentos

Agradeço a Deus, em primeiro lugar, por todas as oportunidades que tive na vida, dentre as quais, a oportunidade de continuar os estudos.

Agradeço aos meus pais que sempre me incentivam e salientam o real valor dos estudos.

Agradeço a minha esposa pela paciência, incentivo e compreensão pelas minhas faltas resultantes desse momento, ao meu filho, que veio nesse mesmo período para renovar as minhas energias.

Agradeço ao meu orientador pela disponibilidade, confiança e por me guiar nesta pesquisa de forma que eu conseguisse chegar até o final e aos meus colegas de turma pela troca de experiências durante o Mestrado.

"Bem-aventurado o homem que acha sabedoria, e o homem que adquire conhecimento; (Bíblia Sagrada, Provérbios 3:13)

"Pois o Senhor é quem dá sabedoria; de sua boca procedem o conhecimento e o discernimento. (Bíblia Sagrada, Provérbios 2:06)

"Pois a sabedoria entrará em seu coração, e o conhecimento será agradável à sua alma. (Bíblia Sagrada, Provérbios 2:10)

Resumo

A Internet possui inúmeros tipos de documentos e é uma influente fonte de informação. O conteúdo Web é projetado para os seres humanos interpretarem e não para as máquinas. Os sistemas de busca tradicionais são imprecisos na recuperação de informações. O governo utiliza e disponibiliza documentos na Web para que os cidadãos e seus próprios setores organizacionais os utilizem, porém carece de ferramentas que apoiem na tarefa da recuperação desses documentos. Como exemplo, podemos citar a Plataforma de Currículos Lattes administrada pelo Cnpq.

A Web semântica possui a finalidade de otimizar a recuperação dos documentos, onde esses recebem significados, permitindo que tanto as pessoas quanto as máquinas possam compreender o significado de uma informação. A falta de semântica em nossos documentos, resultam em pesquisas ineficazes, com informações divergentes e ambíguas. A anotação semântica é o caminho para promover a semântica em documentos.

O objetivo da dissertação é montar um arcabouço com os conceitos da Web Semântica que possibilite anotar automaticamente o Currículo Lattes por meio de bases de dados abertas (Linked Open Data), as quais armazenam o significado de termos e expressões. O problema da pesquisa está baseado em saber quais são os conceitos associados à Web Semântica que podem contribuir para a Anotação Semântica Automática do Currículo Lattes utilizando o Linked Open Data (LOD)?

Na Revisão Sistemática da Literatura foi apresentado conceitos (anotação manual, automática, semi-automática, anotação intrusiva...), ferramentas (Extrator de Entidade...) e tecnologias (RDF, RDFa, SPARQL..) relativas ao tema. A aplicação desses conceitos oportunizou a criação do Sistema Lattes Web Semântico. O sistema possibilita a importação do currículo XML da Plataforma Lattes, efetua a anotação automática dos dados disponibilizados utilizando as bases de dados abertas e possibilita efetuar consultas semânticas.

A validação do sistema é realizada com a apresentação de currículos anotados e a realização de consultas utilizando dados externos pertencentes ao LOD. Por fim é apresentado as conclusões, dificuldades encontradas e proposta de trabalhos futuros.

Palavras-chaves: Anotação Semântica. Dados Abertos Interligados. Plataforma Lattes.

Abstract

The internet presents many different types of documents and is an influential information source. The web content is designed so that human beings are able to understand them, and not the machines. The common used search systems available are imprecise on information recovery. Government uses and make available documents on the Web in order to citizens and their own organization departments use them, but there is a lack of tools to support their tasks of recovery for these documents. For example, the Lattes Platform, provided by CNPq.

The Semantic Web has the purpose of optimizing document recovery, where these documents received synonyms, allowing people and machines to understand the meaning of one information. The lack of semantic in our documents results in ineffective searches that present diverging and ambiguous information. The semantic annotation is the path to promote the semantic in documents.

This dissertation has as objective to build an outline with the Semantic Web concepts that allow to automatically annotate the Lattes Curriculum based on Linked Open Data (LOD), who store terms and expressions' meaning. The problem addressed in this research is based on what of the Semantic Web concepts can contribute to the Automatic Semantic Annotation of the Lattes Curriculum using Linked Open Data.

During the Systematic Review were presented the concepts (manual, automatic, half-automatic, intrusive annotations), tools (Entity Extractor) and technologies (RDF, RDFa, SPARQL...) related to the theme. The application of this concepts allowed the creation of the Semantic Web Lattes System. The system allows importing the XML curricula in the Lattes Platform, annotates automatically the available data using the open databases and allows to run semantic queries.

The system is evaluated by the presenting annotated curricula and by running queries using external LOD data. Finally the conclusions, the drawbacks found and the proposal of future works are presented.

Key-words: Semantic Annotation. Linked Open Data. Lattes Platform.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

Cnpq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
FUMEC	Fundação Mineira de Educação e Cultura
IE	Extração de Informação
LattesWS	Sistema Lattes Web Semântico
LOD	*Linked Open Data*
NER	Reconhecimento de Entidades Nomeadas
NLP	*Natural language processing*
OA	*Open Annotation*
OWL	*Web Ontology Language*
PAS	Plataforma de Anotação Semântica
PLN	Processamento de Linguagem Natural
PRSL	Protocolo de Revisão Sistemática da Literatura
RDF	*Resource Definition Framework*
RSL	Revisão Sistemática da Literatura
SIGC	Sistema de Informação e Gestão do Conhecimento
SPARQL	*SPARQL Protocol and RDF Query Language*
UML	*Unified Modeling Language*
URL	*Uniform Resource Locators*
URI	*Internationalized Resource Identifier*
WS	Web Semântica
W3C	*World Wide Web Consortium*

Sumário

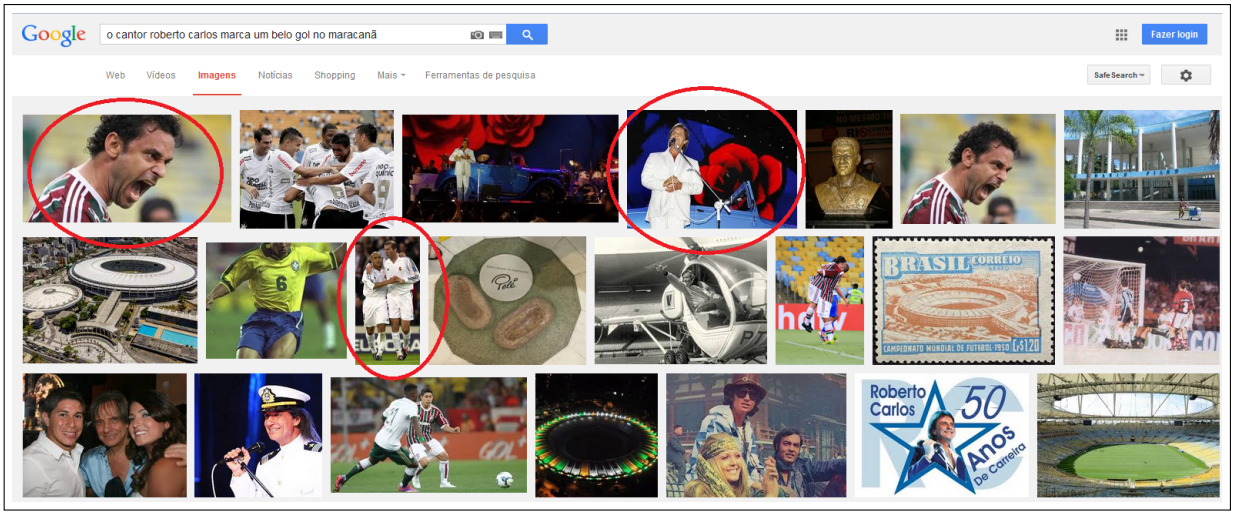
1 Introdução

1.1 Contextualização do Tema

A Internet possui inúmeros tipos de documentos, notícias e informações que podem ser acessadas de qualquer lugar e momento. O conteúdo Web possui pouca ou nenhuma estruturação, com uma formatação HTML, onde parte desse conteúdo é projetada para os seres humanos interpretarem, não para programas de computador manipular. Isso produz nos sistemas de busca tradicionais (utilizam links e palavras chaves) imprecisão e demora na recuperação de informações, isso é, dificultando a realização de pesquisas e retorno de informações incorretas.

A figura de número 1 mostra o conjunto de imagens retornadas no Google referente a pesquisa "o cantor roberto carlos marca um belo gol no maracanã". Para as pessoas é notável que estamos nos referindo ao cantor Roberto Carlos, mas para a máquina de busca isso não está definido, ocasionando resultados irrelevantes.

Figura 1 – Resultados da pesquisa no Google sobre Roberto Carlos e seu show no Maracanã



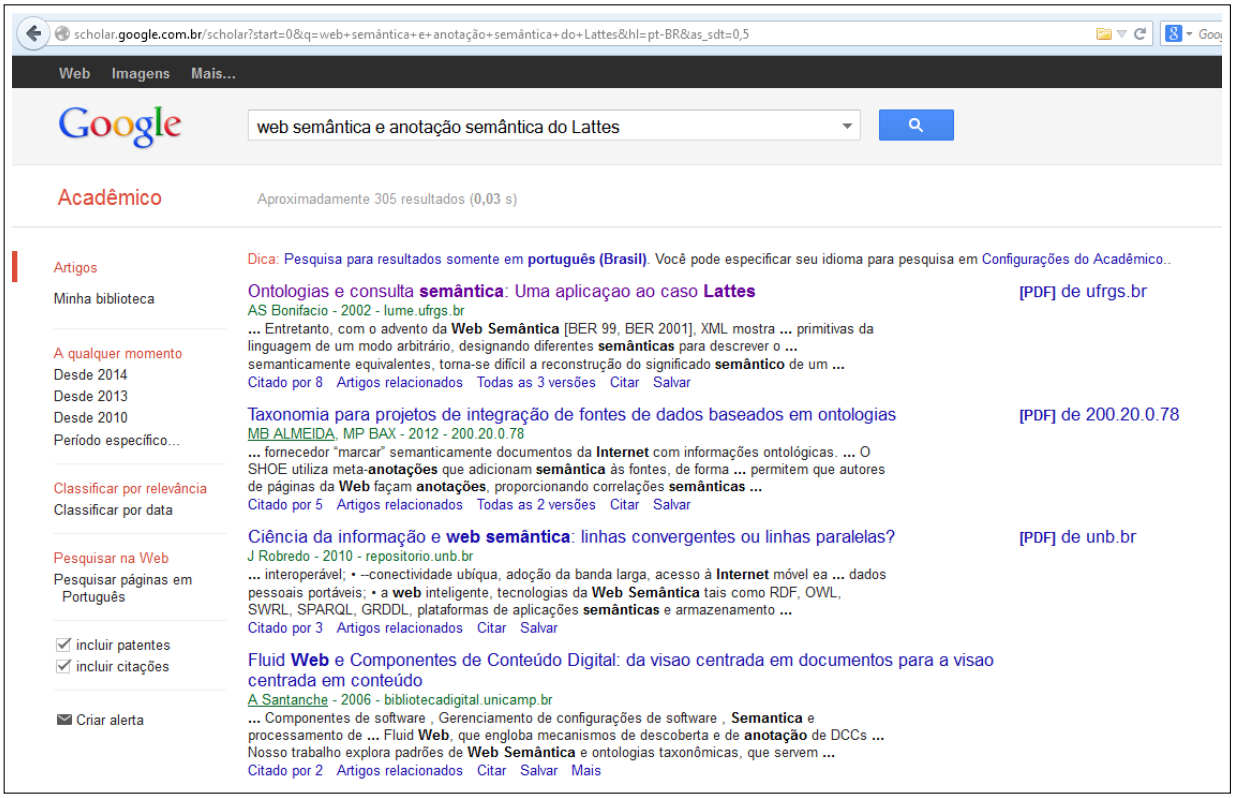
Fonte: Próprio Autor.

O governo utiliza e disponibiliza sistemas de informações na Web para que os cidadãos e seus próprios setores organizacionais utilizem de suas informações e tantas outras disponíveis na Internet, tais como gráficos estatísticos, legislações e documentos oficiais. Com isso é fundamental o auxilio de ferramentas e tecnologias que apoiem na tarefa de recuperar documentos para as pessoas e para os setores do próprio Estado com a finalidade de obter resultados atualizados e automaticamente, agregar valor a projetos e políticas públicas. Recuperar informações no ambiente Web é uma tarefa trabalhosa, principalmente

quando a mesma envolve características específicas em um domínio (área)(FONTES; CAVALCANTI; MOURA, 2013).

A imagem 2 refere-se ao conjunto de resultados do Google Acadêmico, em que foi realizada a pesquisa "web semântica e anotação semântica do Lattes". Retornando 305 páginas que o usuário deverá navegar em busca do documento que venham introduzir o assunto relacionado ao texto pesquisado, a quantidade e a imprecisão nos resultados ocorre porque os motores de busca utilizam as palavras chaves dos documentos e comparam com as palavras do texto da pesquisa, trazendo documentos que tratam de um assunto, em outro momento trazendo documentos relacionados a outra palavra chave, e ainda, documentos que não tem ligação com a pesquisa. De qualquer maneira, essa avaliação é outra tarefa para o usuário. Esse está interessado em encontrar informação onde a relevância dos documentos não podem ser medidas utilizando do uso de sistemas de busca por palavras chaves (Keywords) (BONIFACIO, 2002).

Figura 2 – Resultados da pesquisa no Google sobre web semântica e anotação semântica do Lattes



Fonte: Próprio Autor.

Atualmente pode-se recorrer à Plataforma Lattes, que é a base de dados de currículos mantida e administrada pelo CNPq. Ela disponibiliza dados e informações de pesquisadores, suas experiências, conhecimentos e atividades que estão disponíveis para serem extraídas e relacionadas. Porém, temos poucas consultas disponíveis no site da

instituição para os cidadãos e outros setores organizacionais, sem contar que, o acesso à informação está disponível em um formato sintático, XML representado na figura 3, tornando o processo de leitura e interpretação das informações pelas pessoas incompreensível.

Figura 3 – Exemplo de arquivo XML disponibilizado na extração de dados na Plataforma Lattes CNPq


```
−<CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" DATA-ATUALIZACAO="17062015" HORA-ATUALIZACAO="173348" NUMERO-
IDENTIFICADOR="3564597309576489">
 −<DADOS-GERAIS NOME-COMPLETO="Fernando Silva Parreiras" NOME-EM-CITACOES-BIBLIOGRAFICAS="Parreiras, Fernando Silva;Silva Parreiras, Fernando;PARREIRAS, F.
  S." NACIONALIDADE="B" PAIS-DE-NASCIMENTO="Brasil" UF-NASCIMENTO="MG" CIDADE-NASCIMENTO="Belo Horizonte" PERMISSAO-DE-DIVULGACAO="NAO"
  DATA-FALECIMENTO="" SIGLA-PAIS-NACIONALIDADE="BRA" PAIS-DE-NACIONALIDADE="Brasil">
    <RESUMO-CV TEXTO-RESUMO-CV-RH="<span itemscope itemtype="http://schema.org/Person"><span itemprop="name">Fernando Silva Parreiras</span> possui estágio pós-doutoral
    na <a href="http://www.inf.puc-rio.br/" itemprop="alumniOf">PUC Rio</a> (bolsa <a href="http://cordis.europa.eu/projects/rcn/95951_en.html">Net2 EU FP7 PEOPLE</a>), doutorado em
    Ciência da Computação <a href="http://en.wikipedia.org/wiki/Latin_honors" itemprop="award">Summa Cum Laude</a> pela <a href="http://www.uni-koblenz-landau.de"
    itemprop="alumniOf">Universität Koblenz-Landau</a> na Alemanha (bolsa CAPES/DAAD), mestrado em Ciência da Informação pela <a href="http://eci.ufmg.br/"
    itemprop="alumniOf">UFMG</a>, especialização em Gestão Estratégica pela <a href="http://www.face.ufmg.br/" itemprop="alumniOf">UFMG</a> e graduação em Ciência da Computação
    pela <a href="http://www.fumec.br/" itemprop="alumniOf">FUMEC</a>. Tem experiência no Brasil e no exterior em projetos de pesquisa e desenvolvimento. Na sua carreira acadêmica,
    produziu mais de <a href="#ProducoesCientificas">50 trabalhos completos publicados em periódicos e anais de congressos nacionais e internacionais</a>, <a href="http://scholar.google.com.br
    /citations?user=FQjTCfwAAAAJ"> com quase 600 citações</a>, e o livro <a href="http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118004175.html"
    itemprop="makesOffer">Semantic Web and Software Engineering</a>, Wiley/IEEE. Desde 2011, é <span itemprop="jobTitle">professor e coordenador</span> do <a href="http://ppg.fumec.br
    /sigc/" itemprop="memberOf">Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento</a> da <span itemprop="worksFor">Universidade FUMEC</span>. Foi
    responsável pela intermediação de acordos internacionais de cooperação acadêmica, promovendo a internacionalização do curso. Na pesquisa científica, participa da execução e coordenação de
    <a href="#ProjetosPesquisa" itemprop="seeks"> projetos de P&D no setor elétrico</a>, saúde e engenharia de software, com financiamento <span itemprop="memberOf">FUMEC, FAPEMIG
    e CNPQ</span>. Orienta <a href="#Orientacoesconcluidas">dissertações sobre os temas <span itemprop="seeks">engenharia de software, banco de dados e inteligência analítica<span></a>.
    Na Universidade de Koblenz-Landau, Alemanha, foi o líder do grupo de trabalho WP1 do projeto <a href="http://cordis.europa.eu/projects/rcn/85351_en.html">MOST (Marrying Ontologies
    and Software Technologies)</a>, financiado pelo <a href="http://cordis.europa.eu/fp7">EU FP7-ICT</a>. No escopo dos trabalhos do grupo, foi desenvolvida tecnologia com a <a
    href="http://www.sap.com">SAP Research</a> para validação de modelos de processos de negócio em ferramentas SAP. Desenvolveu tecnologia para modelagem de equipamentos de
    telecomunicações, embarcada em um produto da empresa polonesa <a href="http://www.comarch.com/">Comarch</a>. No ensino, possui 10+ anos de experiência docente em instituições como
    FUMEC, PUC Minas, Uni-BH. Na indústria, é diretor executivo da <a href="http://liaise.com.br/" itemprop="worksFor" itemprop="brand">LIAISE</a>, e atua desde 1999 no setor de
    software: desenvolveu um conjunto de ferramentas para suporte ao desenvolvimento dirigido por modelos (<a href="#SoftwareSemPatente" itemprop="owns">TwoUse Toolkit</a>); atuou no
    desenvolvimento de uma solução de Gestão Eletrônica de Documentos (GED) chamada Docman (Spread); melhorou o desempenho do sistema e diminuiu o tempo gasto com manutenções (attps
    Informática); projetou e desenvolveu um sistema para acesso e controle dos registros da qualidade ISO 9001 (SIARQ) (DATAMEC, Unisys Brasil).</span><img src="https://www.linkedin.com
    /profile/view?id=980240&trk=tab_pro"/>" TEXTO-RESUMO-CV-RH-EN="bachelor's at Ciência da Computação from Universidade FUMEC (2001), master's at Information Science from
    Universidade Federal de Minas Gerais (2005) and doctorate at Ciência da Computação from Universität Koblenz-Landau (2010). Has experience in Computer Science, focusing on Software
    Engeneering, acting on the following subjects: gestão de conteúdo, gestão do conhecimento, ciência da informação, redes sociais and redes de coautoria."/>
    <OUTRAS-INFORMACOES-RELEVANTES OUTRAS-INFORMACOES-RELEVANTES=""/>
 −<ENDERECO FLAG-DE-PREFERENCIA="ENDERECO_INSTITUCIONAL">
    <ENDERECO-PROFISSIONAL CODIGO-INSTITUICAO-EMPRESA="173800000007" NOME-INSTITUICAO-EMPRESA="Universidade FUMEC" CODIGO-
    UNIDADE="" NOME-UNIDADE="" CODIGO-ORGAO="173803000008" NOME-ORGAO="Faculdade de Ciências Econômicas, Administrativas e Contábeis" PAIS="Brasil"
    UF="MG" LOGRADOURO-COMPLEMENTO="Avenida Afonso Pena, 3880" BAIRRO="Cruzeiro" CIDADE="Belo Horizonte" CAIXA-POSTAL="" CEP="30130009" DDD="31"
    TELEFONE="32695230" RAMAL="" FAX="" HOME-PAGE="http://www.fumec.br/"/>
  </ENDERECO>
```


Fonte: Próprio Autor.

A Web Semântica (WS), proposta em 2001 por (BERNERS-LEE; HENDLER; LASSILA, 2001), como uma extensão da Web atual, onde as informações possuem significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação. A Web semântica surge com o propósito de solucionar o problema de recuperação de dados, interoperabilidade e compartilhamento de conhecimento, em que as informações é atribuída (anotada) seus significados, permitindo que tanto as pessoas quanto as máquinas, possam compreender o significado de uma informação. É importante que o domínio de conhecimento a ser partilhado seja descrito de forma genérica e rica, por meio de taxonomias e vocabulários específicos, envolvendo conceitos, propriedades e regras de domínio(FONTES; CAVALCANTI; MOURA, 2013). Essa descrição conceitual, denominada de ontologia, pode ser criada a partir de linguagens lógicas ou ontológicas, como exemplo do RDF (Resouce Description Framework) e OWL (Ontology Web Language), que vai permitir deduzir novas informações sobre um conhecimento, favorecendo a recuperação de informações por agentes inteligentes.

Com a web semântica a Internet passará a funcionar de outra forma, pois em uma rede de informações, cada item passa a conter o seu significado, o que permite melhores interações com o usuário. Diferente da web tradicional, onde os documentos se relacionam utilizando links sem significado definido, essa proposta transmite as palavras significados que viabilizam sistemas de buscas precisos. Assim, não será necessário procurar uma determinada informação em uma série de páginas de resultados genéricos, será exibido páginas que definem a palavra procurada.

1.2 Problema

Com a ausência da semântica em nossos documentos da Internet, temos que principalmente os resultados das pesquisas tornam-se imprecisos, com informações desencontradas e ambíguas. Os motores de buscas por palavra chave não compreendem o real significado da intenção da pesquisa, retornando muitos resultados, ficando a tarefa da compreensão dos resultados e a seleção dos documentos para os usuários. Diante de várias áreas científicas que merecem a atenção na resolução deste tipo de problema, escolhe-se trabalhar com uma que esteja relacionada com a área social e econômica do Brasil.

Quais são os conceitos associados à Web Semântica que podem contribuir para a Anotação Semântica Automática do Currículo Lattes utilizando o Linked Open Data[1]?

1.3 Justificativa

O desenvolvimento da Internet vive um momento em que conceder significado às palavras, aos conteúdos disponíveis nesse ambiente, é importante para o seu avanço e crescimento. A Web Semântica estabelece padrões tecnológicos e ferramentas que possibilitarão a criação de novos ambientes informacionais e a efetivação da Web 3.0. Pesquisas na área da anotação semântica são relevantes para solucionar problemas de busca, de localização e de recuperação da informação.

No Brasil, a Plataforma Lattes, representa a integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um Sistema de Informações. Sua dimensão atual se estende às ações de planejamento, gestão e operacionalização do fomento do CNPq, mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia, das instituições de ensino superior e dos institutos de pesquisa. Além disso, se tornou estratégica para as atividades de planejamento, gestão, para a formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação (CNPQ, 2014).

[1] Dados Abertos Interligado - é um esforço mundial para publicar dados, torna-los abertos e disponíveis para todos usarem.

O Lattes é significativo para as Instituições de Ensino, pois o processamento de seus dados e cadastros permite uma visão e avaliação curricular dos docentes e discentes contemplando os seguintes pontos:

1. Estabelecer uma imagem institucional nos sensos;
2. Formação de grupos de trabalho e pesquisa;
3. Avaliar trabalhos de pesquisadores;
4. Diagnosticar o perfil do pesquisador com outros dentro de sua área de atuação;
5. Controle da produção acadêmica docente e discente;
6. Captação de recursos do Estado e agências de fomento;

1.4 Objetivos

1.4.1 Objetivo Geral

Este trabalho tem como objetivo propor um arcabouço com os conceitos da Web Semântica para anotar automaticamente o Currículo Lattes por meio das ligações de bases abertas (Linked Open Data). O trabalho apresentará conceitos, ferramentas, tecnologias que venham enriquecer os dados do Lattes semanticamente e anotar automaticamente por intermédio dos dados interligados. O propósito é permitir que os dados sejam encontrados e indexados na Web, aberto e disponível em formato compreensível por máquina, além de estar disponível para ser reaplicado em outros sistemas e domínios (GOV, 2014).

1.4.2 Objetivos Específicos

Essa dissertação tem como objetivos específicos:

1. Conceituar e identificar as tecnologias relacionadas com Anotação Semântica e Linked Open Data (LOD).
2. Selecionar os currículos dos docentes da plataforma Lattes/Cnpq.
3. Atribuir significado ao conteúdo do Currículo Lattes com as bases de dados abertas (Anotar o currículo Lattes).

2 Aderência ao Programa de Pós-Graduação em SIGC da Universidade Fumec

O Curso de Mestrado Profissional em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC tem como objetivo geral a geração de conhecimentos e a formação de profissionais mestres com habilidades para o desenvolvimento científico, a produção e aplicação prática de conhecimento no campo interdisciplinar de Sistemas de Informação e Gestão do Conhecimento.

Esse curso de pós-graduação stricto senso é organizado sob a área de concentração de Gestão de Sistemas de Informação e do Conhecimento, e possui como linhas de pesquisas: a linha de Tecnologia e Sistema de Informação e a de Gestão da Informação e do Conhecimento.

Essa dissertação desenvolve-se na linha de Tecnologia e Sistema de Informação porque tem como objetivo definir um conjunto de recursos e soluções da computação que permitam o uso da informação. A finalidade dessa linha é a investigação científica relacionada aos processos que podem favorecer o uso da tecnologia no apoio ao gerenciamento dos sistemas de informação.

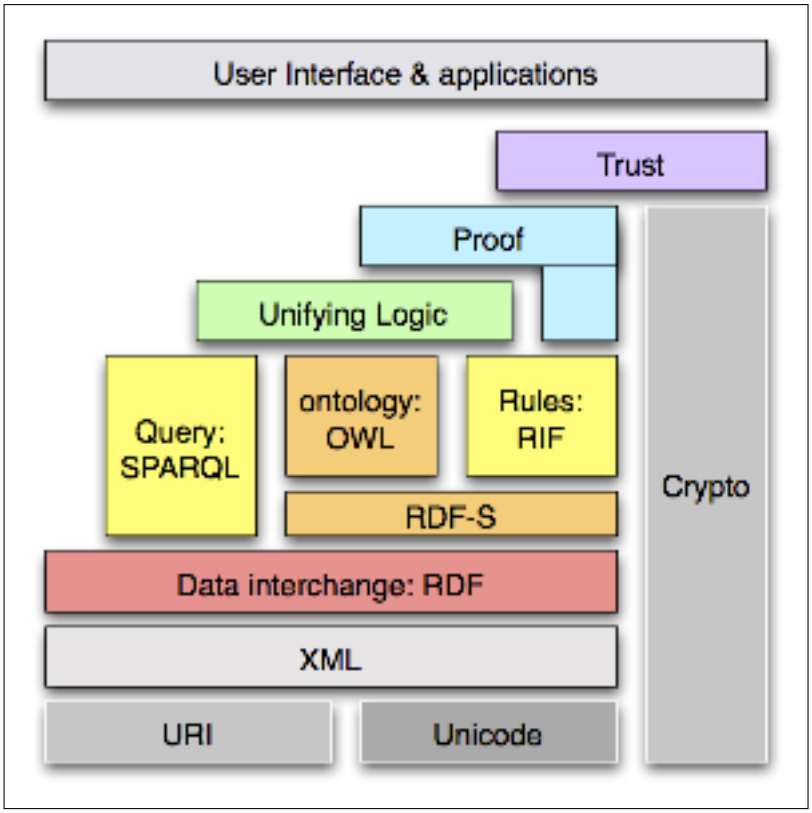
Por fim esse trabalho apresenta uma temática interdisciplinar, pois envolve disciplinas e conceitos de Gestão do Conhecimento, Sistemas e Tecnologias da Informação conforme identificado na revisão sistemática.

3 RSL - Revisão Sistemática da Literatura

3.1 Fundamento e Conceitos da Web Semântica

De acordo com proposta de arquitetura apresentada na figura 4, temos as seguintes tecnologias e camadas relacionada ao projeto Web Semântica:

Figura 4 – Arquitetura da Web Semântica.



Fonte: (BRASIL, 2014)

IRI (International Resource Identifier): é o Identificador Único de Recursos que permite a definição e adoção, de maneira precisa, de nomes aos recursos e seus endereços na Web. É um padrão para identificar um recurso físico ou abstrato de maneira única e global.

UNICODE: é um padrão de codificação dos caracteres, que diminui consideravelmente a possibilidade de redundâncias dos dados, pois funciona independentemente da plataforma utilizada. Ele fornece uma representação numérica universal e sem ambiguidade para cada caractere de maneira independente da plataforma de software e do idioma.

XML (eXtensible Markup Language): é uma linguagem recomendada pela W3C

que permite a criação de documentos que possuem dados estruturados. É uma linguagem que permite a organização dos dados por meio da definição de elementos e atributos, possibilitando através de regras sintáticas análise e validação de recursos. Fornece a interoperabilidade em relação à sintaxe de descrição de recursos da Web Semântica.

RDF (Resource Description Framework): é um modelo de dados que cria declarações no formato de triplas (sujeito, predicado, objeto), possibilitando a descrição dos recursos por meio de suas propriedades e valores.

RDF Schema: é uma extensão da RDF que permite a definição de esquemas para os vocabulários (termos) utilizados nas declarações. É uma linguagem que permite a construção de ontologias com expressividade e inferência, pois fornece um conjunto básico de elementos para a modelagem, e poucos desses elementos podem ser utilizados para inferência.

Ontology/OWL(Web Ontology Language): é uma extensão da RDFS que possibilita a inclusão de elementos com maior poder com relação a expressividade e inferência. É uma linguagem para definir e instanciar ontologias na Web. Ela permite a criação de construtos avançados para descrever semântica de declarações RDFS. É baseada em lógicas descritivas atribuindo poder de raciocínio para a Web Semântica.

SPARQL (Protocol and RDF Query Language): uma linguagem de consulta e protocolo de acesso a dados em RDF. Utilizada na recuperação de informações em aplicações da Web Semântica.

Regras/RIF (Regra Interchange Format): é a camada de suporte a regras, RIF é o formato de regras padrão. Utilizada para descrever relações que não podem ser descritas diretamente com OWL. Ela define regras lógicas relacionadas aos recursos informacionais, possibilitando uma espécie de "Introdução Lógica".

Unifying Logic: camada superior que possibilita a incorporação de "Lógicas Avançadas", isso é, responsável pelo raciocínio e inferência a partir de semântica.

Proof: camada responsável por testar a camada de regras e validar as informações. Ela possibilitará a verificação/comprovação da coerência lógica dos recursos, de modo que os aspectos semânticos das informações estejam descritos de maneira considerável, atendendo a todos os requisitos das camadas inferiores.

Trust: é a camada de confiança, local onde se espera garantir que as informações estejam corretas e confiáveis. Camada onde após serem concluídas as informações das camadas anteriores, se determina uma autenticação para que esses dados tornem-se confiáveis.

Interface: é a última camada que cumpri a interação entre as pessoas e a Web Semântica.

3.1.1 XML - EXtensible Markup Language

A XML é uma linguagem de marcação que trabalha com tags, porém são tags personalizadas (usuários definem suas próprias tags) que permitem a organização e estruturação de dados existentes, possibilitando criar uma marcação para qualquer tipo de informação. O objetivo dessa linguagem é descrever informações (foco nos dados). Essa capacidade de descrição é extremamente importante para o armazenamento, recuperação e transmissão dos dados que são estruturados e compartilhados em distintos sistemas de informação (interoperabilidade de dados).

A figura 5 representa um trecho de um documento XML com informações relativa a professor. Os dados estão estruturados de forma que sabemos que o ID "12345" é referente a um professor e seu nome é "Fernando Wagner". O objetivo é estruturar os dados de maneira que os mesmos estejam em condições de sofrerem processamento. Salienta-se que a sintaxe XML para ser escrita, deve seguir restrições ou regras.

Figura 5 – Trecho de um código XML

```
<---XML--->


<Professor>
  <id>12345</id>
  <Nome>Fernando Wagner</Nome>
  <Area>Banco de Dados</Area>
</Professor>
```

Um documento XML pode expressar um conjunto de dados de diferentes maneiras. A figura 6 representa outra forma no qual os dados do professor poderiam ser dispostos no documento XML. Essa característica pode causar divergência de comunicação entre as aplicações e problemas durante o processamento dos dados. Para resolver esse problema são utilizadas linguagens de definição de esquemas que permite especificar como o documento XML deverá ser escrito. Dentre essas linguagens, podemos destacar a DTD e XML Schema.

Figura 6 – Outra maneira de escrever o trecho do código XML da figura 5.

```
<Professor id='12345'>
  <Descricao>
    <Nome>Fernando Wagner</Nome>
    <Area>Banco de Dados</Area>
  </Descricao>
</Professor>
```

A figura 7 representa um esquema que especifica quais elementos e atributos são permitidos em um documento XML e como estes devem estar. O DTD é antigo e restrito

(tipo de dados delimitado a texto), não possuem semântica, a validação é sintática, tornando os documentos limitados. Porém o XML Schema resolve em parte esse problema. Ele é baseado na própria linguagem XML, permitindo o reúso do código e disponibiliza legibilidade na medida que permite a criação de vocabulários extremamente simples e a definição de tipos de dados como inteiro, binário, entre outros. A W3C recomenda desde 2001, a substituição do DTD pelo XML Schema.

Figura 7 – Trecho código XML Schema

```
1   <?xml version="1.0" encoding="UTF-8"?>
2   <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    elementFormDefault="qualified" attributeFormDefault="unqualified">
3   <xs:complexType name="tEndereco">
4      <xs:sequence>
5        <xs:element name="rua" type="xs:string"/>
6        <xs:element name="numero" type="xs:integer"/>
7              <xs:element name="cidade" type="xs:string" minOccurs="1"
         maxOccurs="unbounded"/>
8                 <xs:element name="estado" type="xs:string"
         minOccurs="0" maxOccurs="unbounded"/>
9      </xs:sequence>
10  </xs:complexType>
11  <xs:element name="cliente" type="tEndereco"/>
12  </xs:schema>
```

A Extensible Markup Language, XML, carrega a sintaxe das informações, com representação sintática de recursos de maneira independente de plataforma e mesmo com o XML Schema, a XML conta com uma semântica limitada com poucas soluções para o processamento de vocabulários. Em resumo, os arquivos XML carregam a sintaxe das informações, mas não a semântica.

Com isso as linguagens RDF e RDF Schema surgem como solução dessas limitações, o que possibilita uma semântica relacionada a identificadores.

3.1.2 RDF - Resource Definition Framework

O XML, apesar de ser uma linguagem recomendada pela W3C, ela não possibilita descrever a semântica de uma informação. Com isso o modelo de dados, Resource Description Framework (RDF), foi proposto como uma solução para a limitação da XML. Esse modelo de dados é baseado na linguagem XML de modo a expressar o significado das informações, permitindo que essas sejam analisadas sintaticamente, possibilitando interoperabilidade entre aplicações, disponibilizando o conteúdo de forma semântica e compreensível pelas máquinas. Com isso, o XML e o RDF tornam-se complementares, a primeira define a estrutura e a segunda permite expressar a semântica associada aos dados.

O RDF é um padrão de modelo para a troca de dados na Web, criado para situações onde a informação precisa ser processada por aplicativos, ao invés de ser mostrado para pessoas. Esse modelo de dados possibilita a definição de sentenças sobre um recurso.

O modelo de dados RDF é definido como:

· Recursos;

· Literais;

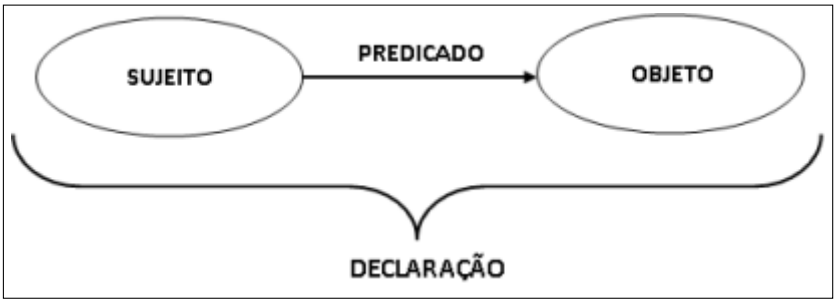
· Propriedades;

· Sentenças.

Pode-se entender que um recurso seja "qualquer coisa" sobre a qual se quer expressar uma idéia. Um recurso pode estar relacionado com dados ou com outros recursos por intermédio das sentenças. Na terminologia da Web, todos os ítens de interesse são chamados de recursos. O recurso é o mapeamento conceitual para uma entidade ou um conjunto de entidades. Ele é um item com uma característica única, especificado por um URI que é um identificador artificial (não transmite qualquer significado). A URI é semelhante a Uniform Resource Locators (URL), porém tanto pode quanto não pode representar uma página Web.

O relacionamento entre um recurso e um literal é chamado de sentença. De acordo com a figura 8, uma sentença é implementada no formato de triplas (frases formadas com sujeito, predicado e objeto). A sentença relaciona um objeto a um sujeito por meio de um predicado. O sujeito é uma URI, o objeto pode ser uma URI ou um texto e o predicado define como o sujeito e o objeto se relacionam. O sujeito e o objeto representam um recurso ou itens de interesse em um domínio (BIZER TOM HEATH, 2009).

Figura 8 – Modelo gráfico de representação RDF



Fonte: Próprio Autor.

Um documento RDF pode ser representado de forma abstrata da seguinte maneira:

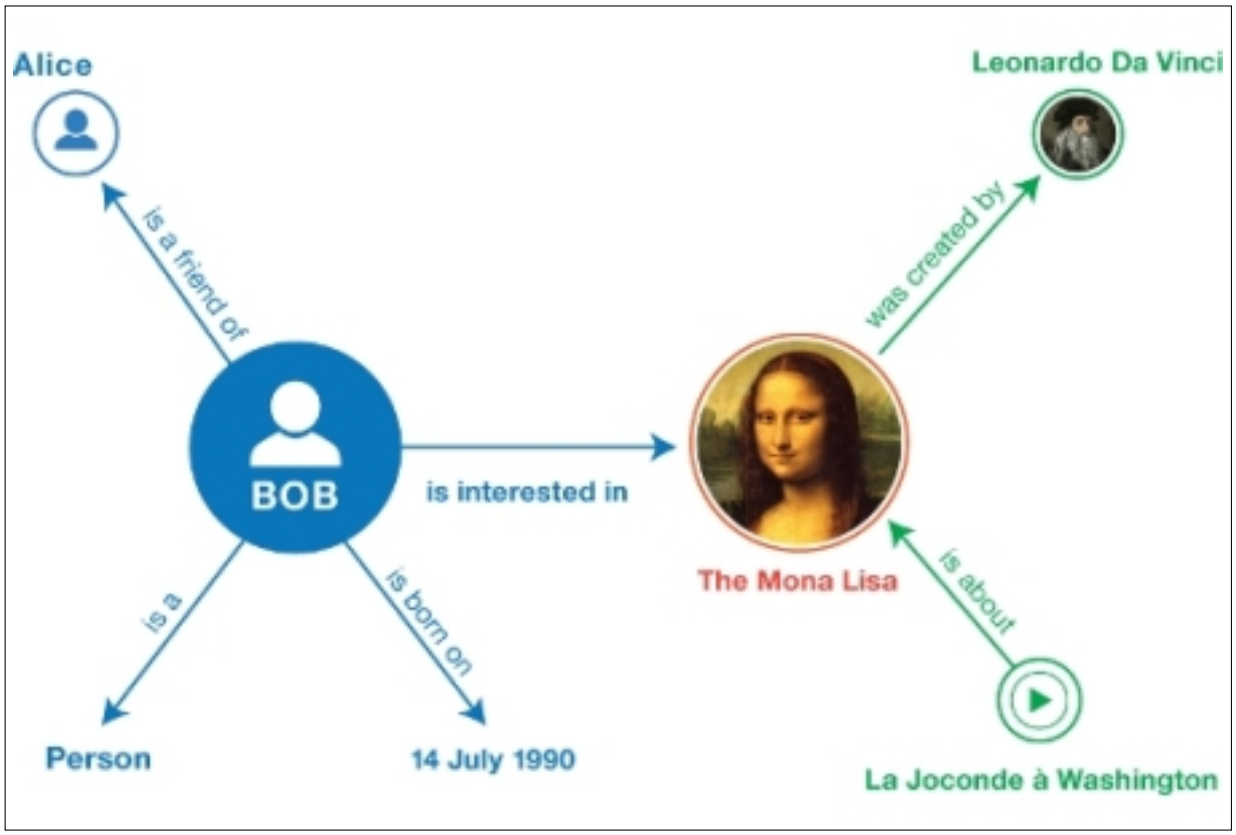
1. <Bob> <is a> <person>.
2. <Bob> <is a friend of> <Alice>.
3. <Bob> <is born on> <the 4th of July 1990>.
4. <Bob> <is interested in> <the Mona Lisa>.

5. <the Mona Lisa> <was created by> <Leonardo da Vinci>.
6. <the video 'La Joconde à Washington'> <is about> <the Mona Lisa>

Um conjunto de triplas que descrevem informações sobre os recursos envolvidos no domínio de interesse.

A figura 9 representa a forma gráfica geral de representar um documento RDF, no qual cada recurso e/ou literais existentes são os nós e as propriedades são as arestas.

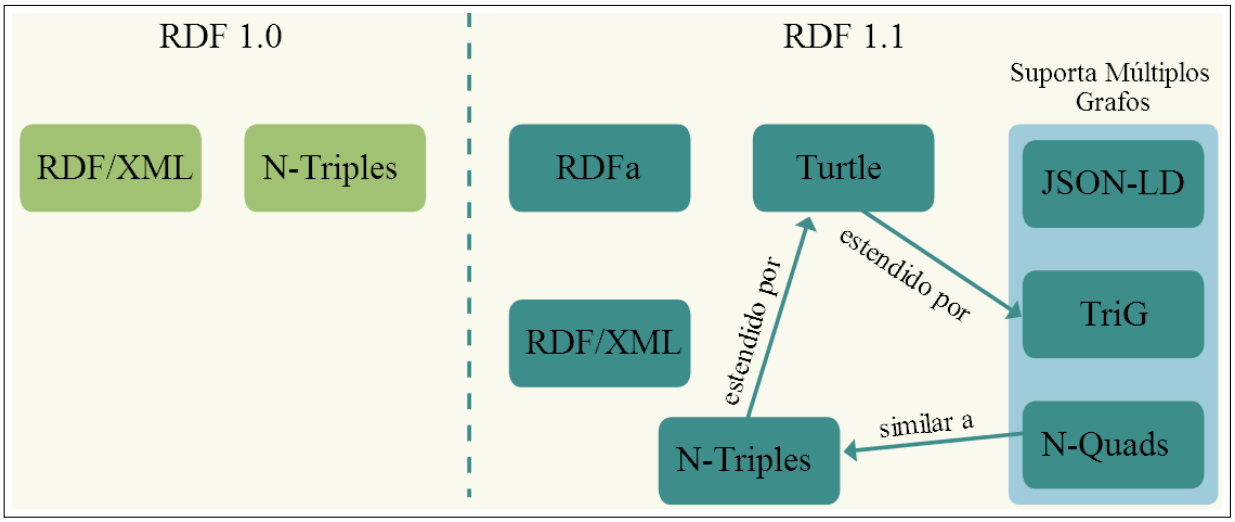
Figura 9 – Representação gráfica de um documento RDF



Fonte: (PRIMER, 2014)

Um aspecto para o RDF é com relação os formatos em que ele pode ser representado. A figura 10 representa a evolução da forma como um Gráfico RDF pode ser escrito/serializado.

Figura 10 – Representação da evolução dos documentos RDF



Fonte:(WEB, 2016)

RDF/XML: Foi a primeira serialização feita para o RDF, possui a sua estrutura dentro do XML e da tag <rdf:RDF>, o elemento rdf:Description é utilizado para identificar sujeitos que é descrito pelo rdf:about. A figura 11 representa triplas descritas utilizando a sintaxe XML, que habilita o intercâmbio entre máquinas, sem interferência humana através de aplicações e serviços.

Figura 11 – Representação de Triplas RDF utilizando a sintaxe XML

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
         xmlns:dcterms="http://purl.org/dc/terms/"
         xmlns:foaf="http://xmlns.com/foaf/0.1/"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:schema="http://schema.org/">
   <rdf:Description rdf:about="http://example.org/bob#me">
      <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
      <schema:birthDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1990-07-04</schema:birthDate>
      <foaf:knows rdf:resource="http://example.org/alice#me"/>
      <foaf:topic_interest rdf:resource="http://www.wikidata.org/entity/Q12418"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://www.wikidata.org/entity/Q12418">
      <dcterms:title>Mona Lisa</dcterms:title>
      <dcterms:creator rdf:resource="http://dbpedia.org/resource/Leonardo_da_Vinci"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619">
      <dcterms:subject rdf:resource="http://www.wikidata.org/entity/Q12418"/>
   </rdf:Description>
</rdf:RDF>
```

Fonte: (PRIMER, 2014)

RDF N3 (N-Triples): Faz parte da família Turtle, sendo o exemplo sucinto de representar a tripla <sujeito> <predicado> <objeto>. Isso é, cada linha de um arquivo

N3 é representado exatamento por uma tripla e no fim da linha um ponto (.) indicando o fim da tripla.

RDFa (Resource Description Framework in Attributes): tem por objetivo embutir código RDF em estruturas HTML e XML realizado através da inclusão de significado via atributos dos elementos. A vantagem da utilização do RDFa é que máquinas de buscas podem melhorar seus resultados aumentando a precisão sobre o real significado de um documento. Ou seja, as máquinas de buscas podem agregar os dados de um documento com dados de outro documento, enriquecendo os resultados de buscas ((PRIMER, 2014),(WEB, 2016)).

Figura 12 – Representação de Triplas RDFa

```
01 <body prefix="foaf: http://xmlns.com/foaf/0.1/
02                schema: http://schema.org/
03                dcterms: http://purl.org/dc/terms/">
04   <div resource="http://example.org/bob#me" typeof="foaf:Person">
05     <p>
06       Bob knows <a property="foaf:knows" href="http://example.org/alice#me">Alice</a>
07       and was born on the <time property="schema:birthDate">1990-07-04</time>.
08     </p>
09     <p>
10       Bob is interested in <span property="foaf:topic_interest"
11       resource="http://www.wikidata.org/entity/Q12418">the Mona Lisa</span>.
12     </p>
13   </div>
14   <div resource="http://www.wikidata.org/entity/Q12418">
15     <p>
16       The <span property="dcterms:title">Mona Lisa</span> was painted by
17       <a property="dcterms:creator" href="http://dbpedia.org/resource/Leonardo_da_Vinci">Leonardo da Vinci</a>
18       and is the subject of the video
19       <a href="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619">'La Joconde à Washington'</a>.
20     </p>
21   </div>
22   <div resource="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619">
23       <link property="dcterms:subject" href="http://www.wikidata.org/entity/Q12418"/>
24   </div>
25 </body>
```

Fonte: (PRIMER, 2014)

Ainda sobre RDFa e de acordo com (WEB, 2016): Na figura 12 observamos que no código HTML aparecem quatro atributos com o objetivo de descrever código RDF:

1. Prefix, que tem por objetivo descrever os vocabulários que estão sendo reusados no documento HTML. Neste caso, temos o reuso de três vocabulários conhecidos por profissionais da área (FOAF, Schema.org e Dublin Core);

2. Resource, cujo objetivo é descrever um determinado recurso. Por exemplo, na linha 4 temos a descrição do recurso <http://example.org/bobme>;

3. Property, cujo objetivo é descrever uma propriedade. Semelhante ao atributo resource, o atributo property é similar à propriedade rdf:Property. Uma propriedade tem o objetivo de relacionar dois elementos, ou seja relacionar um sujeito a um objeto. Isto quer dizer que uma propriedade irá relacionar dois recursos. Podemos observar que como consequência a isto, todas as propriedades apresentadas no código da Figura 12 aparecem dentro da estrutura de um elemento .

Isto quer dizer, por exemplo, que das linhas 4 a 13, há 3 triplas relacionadas à Bob (http://example.org/bobme foaf:knows Alice ; http://example.org/bobme schema:birthDate 1990-07-04 ; http://example.org/bobme foaf:topic$_{i}nteresttheMonaLisa$);

4. Typeof, que equivale ao atributo para representar o elemento rdf:type (tem o mesmo objetivo do rdf:type). Por exemplo, na linha 04 temos: http://example.org/bobme rdf:type foaf:Person.

RDF TURTLE (ttl): Representa uma evolução do N3, sendo criado para possibilitar descrever prefixos e IRIs relativos na estrutura do documento. Com a figura 13 observarmos que as seis primeiras linhas do código mostram os IRIs que podem ser definidos como prefixos e IRI base do documento, característica esta não permitida no N-Triples. Podemos observar que as linhas 08, 14 e 18 apresentam sujeitos com seus predicados e objetos logo abaixo deles. Este tipo de organização e endentação torna intuitiva a leitura do documento, facilitando a identificação das triplas RDF. O elemento que é responsável por relacionar o sujeito ao predicado deste exemplo é o token "a". Este token "a" possui a mesma semântica da propriedade rdf:type e é usada para dizer que bobme é do tipo foaf:Person ((PRIMER, 2014),(WEB, 2016)).

Figura 13 – Representação de Triplas RDF Turtle (ttl)

```
01  BASE   <http://example.org/>
02  PREFIX foaf: <http://xmlns.com/foaf/0.1/>
03  PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
04  PREFIX schema: <http://schema.org/>
05  PREFIX dcterms: <http://purl.org/dc/terms/>
06  PREFIX wd: <http://www.wikidata.org/entity/>
07
08  <bob#me>
09      a foaf:Person ;
10      foaf:knows <alice#me> ;
11      schema:birthDate "1990-07-04"^^xsd:date ;
12      foaf:topic_interest wd:Q12418 .
13
14  wd:Q12418
15      dcterms:title "Mona Lisa" ;
16      dcterms:creator <http://dbpedia.org/resource/Leonardo_da_Vinci> .
17
18  <http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619>
19      dcterms:subject wd:Q12418 .
```

Fonte: (PRIMER, 2014)

RDF TriG: é uma extensão do formato Turtle, isso é, herda as mesmas características. Sua diferença está na possibilidade de representar múltiplos grafos, que foi adicionado a partir da especificação 1.1 do RDF. A diferença entre o código Turtle e o código TriG é que os sujeitos descritos são encapsulados no interior da palavra-chave

chamada GRAPH. Um exemplo de código TriG para representação de múltiplos grafos foi apresentado na Figura 14.

Figura 14 – Representação de Triplas RDF TriG

```
BASE   <http://example.org/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX schema: <http://schema.org/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX wd: <http://www.wikidata.org/entity/>

GRAPH <http://example.org/bob>
  {
    <bob#me>
        a foaf:Person ;
        foaf:knows <alice#me> ;
        schema:birthDate "1990-07-04"^^xsd:date ;
        foaf:topic_interest wd:Q12418 .
  }

GRAPH <https://www.wikidata.org/wiki/Special:EntityData/Q12418>
  {
    wd:Q12418
        dcterms:title "Mona Lisa" ;
        dcterms:creator <http://dbpedia.org/resource/Leonardo_da_Vinci> .

    <http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619>
       dcterms:subject wd:Q12418 .
  }

<http://example.org/bob>
    dcterms:publisher <http://example.org> ;
    dcterms:rights <http://creativecommons.org/licenses/by/3.0/> .
```

Fonte: (PRIMER, 2014)

RDF N-Quads: É uma extensão do N3 e é utilizado para permitir o intercâmbio de catálogo de dados. Ele permite adicionar um quarto elemento a uma linha, capturando o gráfico IRI da tripla descrita na linha.

RDF JSON-LD: Essa é uma extensão da proposta do JSON, com o objetivo de transformar código JSON para RDF com o mínimo de esforço. Este formato é intuitivo para programadores familiarizados com a sintaxe JSON.

3.1.3 RDFs - Resource Definition Framework Schema

O RDFs é um complemento para o RDF com o objetivo de oferecer um suporte para a criação de ontologias. O RDFs é uma extensão semântica do RDF, que fornece maneiras para descrever grupos de recursos e as relações entre esses recursos. Esses recursos são utilizados para especificar as características de outros recursos, como domínios e faixas de propriedades (SCHEMA1.1, 2014).

A ideia é unir RDFs + RDF de tal forma que todas as sentenças descritas em RDF obedeçam à semântica descrita no esquema especificado em RDFs. O RDF Vocabulary Description Language ou RDFs é um vocabulário para descrever classes e propriedades dos objetos baseados em RDF com semântica para hierarquias generalizadas dessas propriedades e classes. O RDFs possibilita trabalhar com relacionamentos de abstração de agregação, generalização ou especialização e associação.

Classes descrevem conceitos de um domínio, possibilitando a modelagem do domínio de interesse. As classes são os próprios recursos. Eles são frequentemente identificados por IRIs e podem ser escritos utilizando propriedades de RDF, que é uma relação entre recursos de sujeito e objeto (SCHEMA1.1, 2014) .

De acordo com o (W3C-RDF1.1-PRIMER, 2014) os construtores que permitem especificar formalmente um esquema está sendo demonstrado na figura 15.

Figura 15 – Construtores da Modelagem RDFs.

Construct	Syntactic form	Description
Class (a class)	**C** `rdf:type rdfs:Class`	**C** (a resource) is an RDF class
Property (a class)	**P** `rdf:type rdf:Property`	**P** (a resource) is an RDF property
type (a property)	**I** `rdf:type` **C**	**I** (a resource) is an instance of **C** (a class)
subClassOf (a property)	**C1** `rdfs:subClassOf` **C2**	**C1** (a class) is a subclass of **C2** (a class)
subPropertyOf (a property)	**P1** `rdfs:subPropertyOf` **P2**	**P1** (a property) is a sub-property of **P2** (a property)
domain (a property)	**P** `rdfs:domain` **C**	domain of **P** (a property) is **C** (a class)
range (a property)	**P** `rdfs:range` **C**	range of **P** (a property) is **C** (a class)

Fonte: (PRIMER, 2014)

A figura 16 é um exemplo de um modelo RDFs: na linha 2, destacamos a utilização de *namespaces* visando distinguir o contexto dos elementos utilizados; nas linhas 3-6 apresentam declarações de duas **classes**, *Pesquisador* e *Evento*, juntamente com uma instância de cada uma destas classes. Nas linhas 8-11, é declarada uma **propriedade** *Envolve* que relacionará instâncias de pesquisadores com instâncias de eventos. E nas linhas 13-15, apresentam a declaração de outra **propriedade**, chamada *Organiza*. Os conjuntos *domain* e *range* desta nova propriedade não estão declarados de maneira explicita, mas pode-se inferir que estes têm como valores os conjuntos de pesquisadores e eventos respectivamente, pelo fato de que a propriedade *Organiza* é uma **subpropriedade** de *Envolve*.

Os links RDF são a base dos dados ligados. Eles permitem que as aplicações cliente naveguem entre as fontes de dados e descubram dados adicionais. Para fazer parte da Web de Dados, fontes de dados devem definir links RDF para relacionar as entidades em outras fontes de dados (BIZER TOM HEATH, 2009).

Figura 16 – Declaração de domains, range e subPropertyOf de um RDFS.

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/#" >
    <rdfs:Class rdf:ID="Pesquisador"/>
    <rdfs:Class rdf:ID="Evento"/>
        <Pesquisador rdf:ID="Walison"/>
        <Evento rdf:ID="SBBD"/>

    <rdf:Property rdf:ID="Envolve">
        <rdfs:domain rdf:resource="#Pesquisador"/>
        <rdfs:range rdf:resource="#Evento"/>
    </rdf:Property>

    <rdf:Property rdf:ID="Organiza">
        <rdfs:subPropertyOf rdf:resource="#Envolve"/>
    </rdf:Property>

```

Fonte: Próprio Autor.

3.1.4 OWL - Web Ontology Language

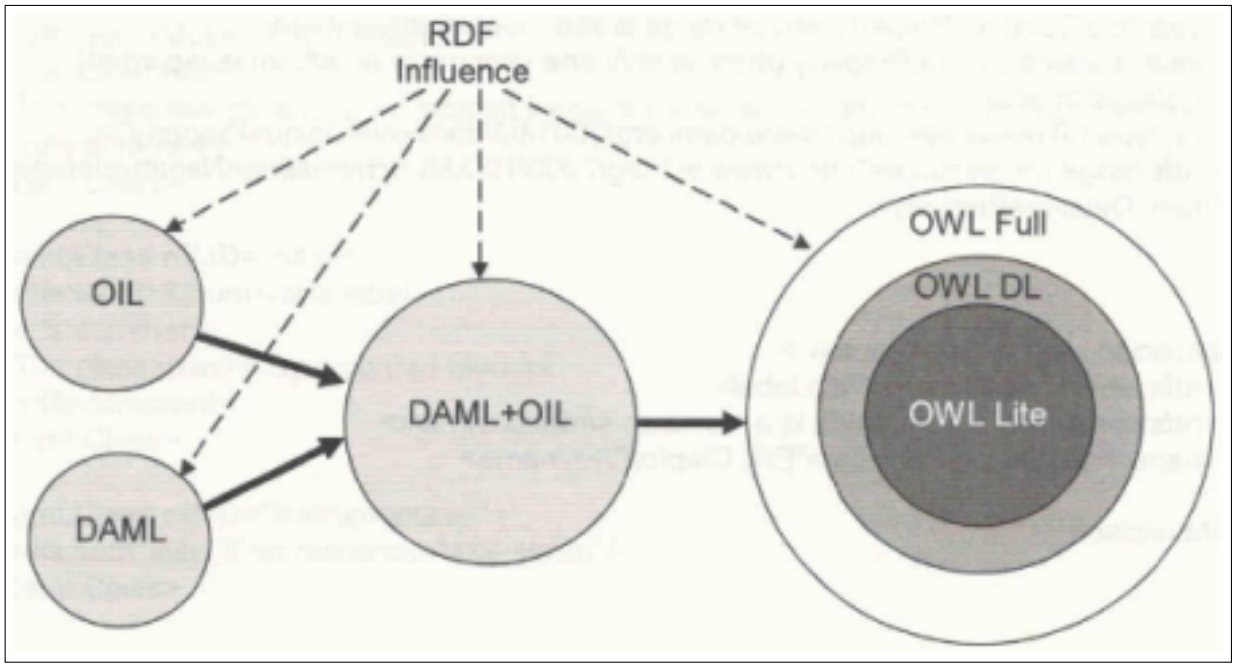
A Linguagem de Ontologia Web (OWL) é utilizada para definir e instanciar ontologias (modelo de dados que representa um conjunto de conceitos de uma área de interesse e os relacionamentos entre essas) na Web com recomendação da W3C. Essa linguagem possibilita que as informações contidas em documentos possam ser interpretadas por pessoas e máquinas, permitindo a interoperabilidade entre aplicações.

Segundo (GALEGO, 2013), OWL foi estabelecido pelo Grupo de Trabalho em Ontologia para Web do W3C como linguagem padrão para construir ontologias para a infraestrutura da Web Semântica. A figura 17 remete a origem da OWL, indicando que ela é o fruto da fusão de duas outras linguagens de descrição: DAML-ONT e OIL. Por isto, inicialmente, era conhecida como DAML-OIL. OWL pode ser entendida como uma extensão do RDF Schema, utilizando dos conceitos anteriormente definidos (como rdfs:Class e rdfs:subClass) para suportar uma plena expressividade.

De acordo com (OWL1, 2014), essa linguagem proporciona expressar ricos e complexos relacionamentos, permitindo, a criação de aplicações com uma inferência ou raciocínio superior, ela agrega as características que existem no RDF e RDFS. As sub-línguas para o desenvolvimento de ontologias que são incrementalmente expressivas são:

OWL Lite é para aqueles usuários que necessitam utilizar uma sintaxe simples e que apoie uma hierarquia de classificação e restrições primárias. Por exemplo, embora suporte restrições de cardinalidade, ela permite valores de cardinalidade 0 ou 1. A vantagem é que essa linguagem tem uma menor complexidade formal que OWL DL, sendo de maior entendimento por parte de usuários e sendo descomplicada na implementação

Figura 17 – Influência do RDF na criação da OWL



pelos desenvolvedores. Ela possui decidibilidade computacional, ou seja, a computação terminará em um tempo finito.

OWL DL é utilizada por usuários que desejam uma expressividade superior a do que a oferecida pelo OWL Lite. Baseia-se em lógica descritiva, um fragmento de Lógica de Primeira Ordem, passível portanto, de raciocínio automático. Possui completude computacional e decidibilidade. Possui restrições, embora uma classe possa ser subclasse de muitas classes, uma classe não pode ser instância de outra classe. OWL DL possui essa denominação por causa da sua correspondência com lógica descritiva, um campo de pesquisa que estuda as lógicas que formam a base formal da OWL.

OWL Full é para aqueles usuários que almejam a expressividade suprema e a liberdade sintática do RDF sem garantia computacional. Como exemplo, uma classe pode ser tratada simultaneamente como uma coleção de indivíduos e como um indivíduo nela própria. Devido a completude da linguagem, é improvável que softwares de inferência venham a ser capazes de suportar por completo cada recurso da OWL Full.

Existe compatibilidade entre as sub-línguas no sentido de que OWL Lite está contido na OWL DL que por sua vez está contida na OWL Full, sendo essa, integralmente compatível com RDF, tanto no sintático quanto no semântico.

Os componentes elementares de uma ontologia OWL são: indivíduos, propriedades e as classes. Estas classes, propriedades e indivíduos podem ser vistos como constituintes atômicos de axiomas e são chamados de entidades.

As **classes** representam os diferentes conceitos, domínios e são as principais enti-

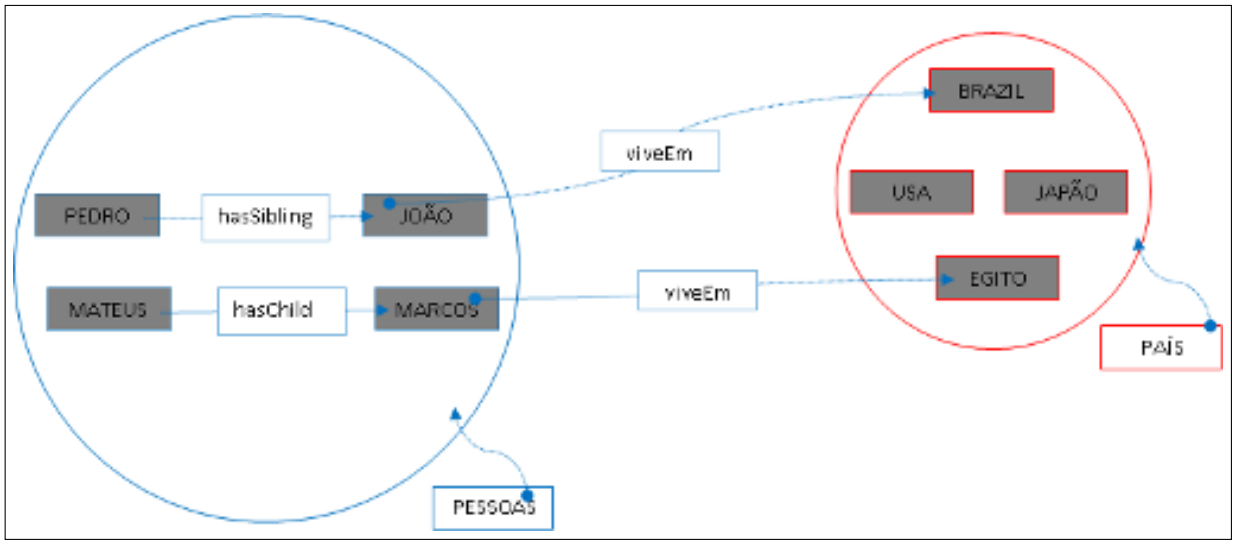
dades de uma ontologia. Podem organizar-se em hierarquias de superclasse e subclasse, conhecidas como taxonomias. A classe é o conjunto que contêm os indivíduos e são construídas a partir de descrições, as quais especificam as condições que devem ser satisfeitas por um indivíduo para que ele possa ser um membro da classe. Subclasses são especializações de suas superclasses.

Propriedades são relações binárias (relações que contém duas coisas) entre indivíduos (instância de uma classe), ou seja, as propriedades ligam dois indivíduos. Na UML é conhecido como relação. Por exemplo, a propriedade hasSibling (temIrmão) pode ligar o indivíduo Pedro ao indivíduo João; ou a propriedade hasChild (temCriança) pode ligar o indivíduo Marcos ao indivíduo Mateus. As Propriedades podem ser inversas. Por exemplo, a propriedade inversa de hasOwner (temDono) é isOwnedBy (éPropriedadeDe). As propriedades podem limitar-se a um único valor: são as Functional Properties (propriedades funcionais). Elas podem ser Transitive Properties (Propriedades transitivas) ou Symetric Properties (Propriedades Simétricas).

Os **indivíduos** são as instâncias das classes de uma ontologia OWL. Os indivíduos herdam as propriedades das classes de que são membros. Em OWL dois nomes diferentes podem remeter ao mesmo indivíduo.

A figura 18 representa um exemplo de relação entre classes, indivíduo e suas propriedades. Com a OWL é possível combinar diferentes classes e ou propriedades para criar outras classes e propriedades. Essas combinações podem ser complexas e é denominada de expressões. As expressões são as principais razões do poder de expressividade da linguagem OWL. O grupo de trabalho OWL do W3C, em 2009, fizeram uma revisão dessa linguagem e adicionou recursos novos, resultando em uma versão atual da OWL conhecida como OWL 2.

Figura 18 – Exemplo de Relação entre classes e indivíduos



Fonte: Próprio Autor.

3.1.5 OWL 2

Em 2009, o W3C estendeu as funcionalidades do OWL, definindo uma linguagem chamada OWL 2, ficando a primeira versão de 2004, definida a partir desse momento como OWL 1. De acordo com (OWL2, 2015) assim como a OWL 1, a OWL 2 é criada para facilitar o desenvolvimento de ontologias e compartilhamento do conhecimento na Web, com o objetivo de expandir a interoperabilidade entre as aplicações de software. A compatibilidade do OWL 2 com OWL 1 é completa: Toda ontologia no formato OWL 1 permanece válida para OWL 2, com inferências idênticas em todos os casos.

Com relação ao OWL 1, a OWL 2 adiciona funcionalidades. De acordo com (GALEGO, 2013), as novas funcionalidades são para harmonizar a forma sintática (por exemplo, a união disjunta de classes), e outros para oferecer uma expressividade diferente:

1. (keys) Definições de chaves para identificar indivíduos de forma exclusiva, por exemplo: CPF de uma pessoa, número da placa e estado para um veículo automotivo.

2. (property chains) Propriedades em cadeia: Provê uma forma de definir propriedades a partir de uma composição de outras propriedades. Por exemplo: para definirmos a propriedade TIO, podemos utilizar as propriedades PAI e IRMÃO.

3. (richer datatypes, data ranges) Tipo de dados compostos, faixa de dados.

4. (qualified cardinality restrictions) Restrições de cardinalidade.

5. (asymmetric, reflexive, and disjoint properties) Propriedades assimétricas, reflexivas e disjuntas.

6. (enhanced annotation capabilities) Estende o uso de anotações, como um comentário ou uma descrição, permitindo aplicar na ontologia, entidades, indivíduos anônimos, axiomas e nas próprias anotações.

Também foi definido novas sub-linguagens ou *profiles*: OWL 2 EL, OWL 2 QL e OWL 2 PL.

OWL 2 EL: Executa algoritmos em tempo polinomial para todas as tarefas padrões de raciocínio. É adequada para aplicações com vasta ontologia, nas quais capacidade de expressão pode ser substituída por garantia de performance.

OWL 2 QL: Retorna consultas conjuntivas em tempo logarítmico usando tecnologia de banco de dados relacional. É adequada para aplicações nas quais ontologias são usadas para organizar um extenso número de indivíduos e nas quais é necessário acessar dados diretamente via consultas relacionais (exemplo: SQL).

OWL 2 RL: Executa algoritmos de raciocínio em tempo polinomial usando tecnologias de banco de dados com regras estendidas diretamente sobre as triplas RDF. É adequado para aplicações que exigem um alto grau de raciocínio (por meio de regras de inferência) sem, no entanto, comprometer o poder de expressividade.

3.1.6 SPARQL - Protocol and Query Language

Na Web Semântica os dados são representados utilizando o modelo de dados conceitual de RDF, acompanhado com as extensões de RDFS e OWL. Dados que estão armazenados em um banco de dados de triplas ou relacional com esquema de mapeamento para RDF. A partir disso os dados estão disponíveis para que seja realizado as manipulações de dados.

SPARQL - Protocol and Query Language: é a linguagem responsável por efetuar as manipulações de dados da Web Semântica. Os resultados de anotações são salvos no formato Resource Description Framework (RDF), que fornece um modo padrão para compartilhamento de dados, intercâmbio, permite consultar e manipular os dados usando a linguagem de consulta SPARQL (TAO et al., 2013). Assim como, a linguagem SQL está para os bancos de dados relacionais, também está o SPARQL para os bancos de dados de triplas RDF. Entretanto, consultas que necessitam relacionar diferentes dados tendem a ser bem complexas em SQL do que em SPARQL (CASTAñO, 2008).

SPARQL possibilita a recuperação de dados estruturados e semi-estruturados, a exploração de dados em consultas com relações desconhecidas e uniões de conjunto de diversos dados em uma consulta. O resultado de uma consulta SPARQL pode ser um

conjunto de resultados ou um grafo RDF (1.1, 2013). Quanto maior for o número de informações relacionadas, maior será a vantagem do SPARQL diante do SQL na facilidade de criação de consultas (CASTAñO, 2008).

A imagem 19 representa a forma geral da descrição de consultas SPARQL. Em PREFIX declaramos os namespaces utilizados na consulta, em SELECT declara-se o conjunto de resultados almejados, em FROM declara-se o conjunto de dados a serem consultados (os grafos RDF que serão consultados), na cláusula WHERE monta-se a condição de triplas que o resultado da pesquisa deverá satisfazer e ORDER BY, DISTINCT, LIMIT, entre outros são os modificadores da consulta.

Figura 19 – Modelo Geral da Estrutura de Consulta em SPARQL

```
PREFIX
Ex.   PREFIX f: <http://example.org#>

SELECT
Ex.   SELECT ?age

FROM
Ex.   FROM <http://www.wssistemas.com.br/latteswss/recursordf/3564597309576489.rdf>

WHERE
Ex.   WHERE { f:mary f:age ?age }

ORDER BY, DISTINCT, etc.
Ex.   ORDER BY ?age
```

Fonte:Próprio Autor.

Exemplo de Consulta: Na figura 20 temos um exemplo de consulta Sparql. Supondo que temos essas triplas armazendas em um banco de dados de triplas, efetuamos essa consulta com o intuito de selecionar o nome de um livro e em seguida demonstramos o resultado dessa pesquisa.

Figura 20 – Exemplo de Consulta em SPARQL

Fonte:(WEB, 2016)

Exemplo de Insert: Com o comando INSERT DATA é possível inserir triplas RDF em repositórios de triplas. A próxima imagem, figura 21, representa a inserção de duas triplas em um Graph padrão de um banco de dados de triplas.

Figura 21 – Exemplo de Insert em SPARQL

Tripla RDF (antes)

```
# Default graph
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix ns: <http://example.org/ns#> .

<http://example/book1> ns:price 42 .
```

Insert de Triplas RDF

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
INSERT DATA
{
  <http://example/book1> dc:title "A new book" ;
                         dc:creator "A.N.Other" .
}
```

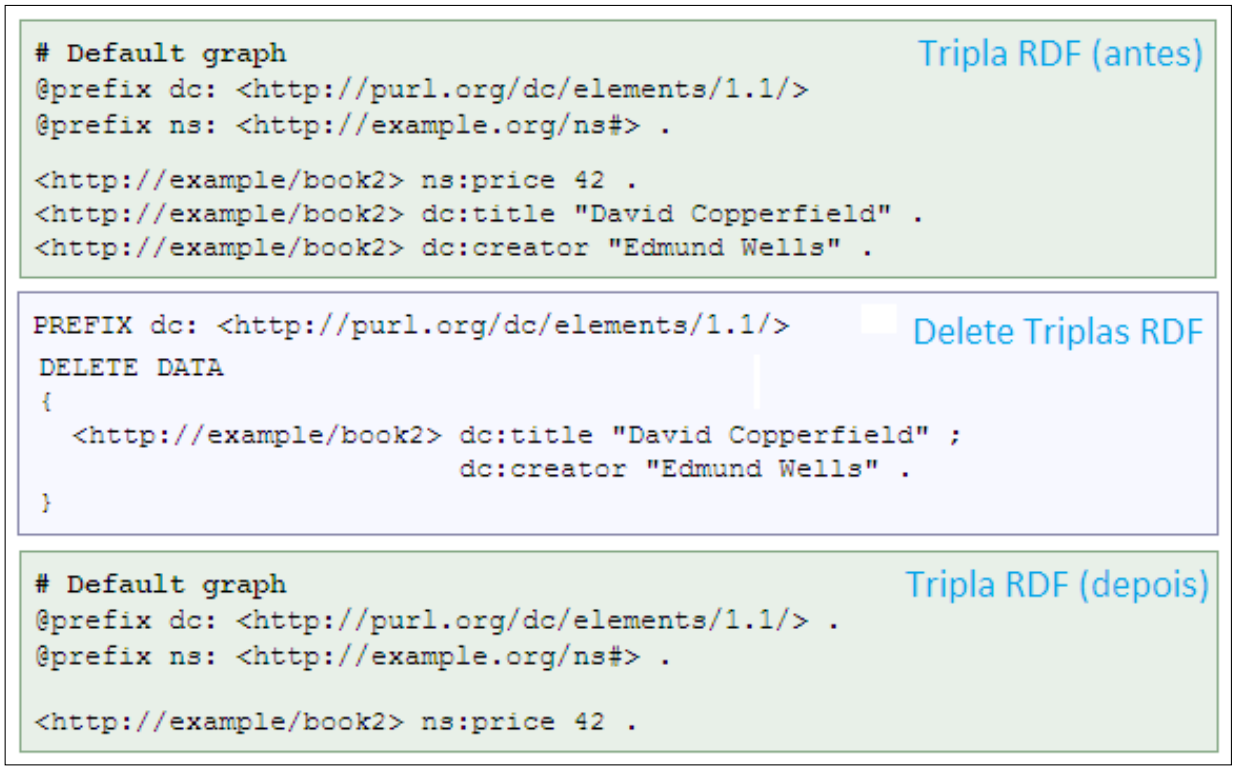
Tripla RDF (depois)

```
# Default graph
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix ns: <http://example.org/ns#> .

<http://example/book1> ns:price 42 .
<http://example/book1> dc:title "A new book" .
<http://example/book1> dc:creator "A.N.Other" .
```

Fonte:(1.1, 2013)

Exemplo de Delete: Esse exemplo da figura 22 representa a remoção de triplas a partir do comando DELETE DATA. Essa figura exemplifica a remoção de duas triplas RDF.

Figura 22 – Exemplo de Delete em SPARQL



Fonte:(1.1, 2013)

As consultas são executadas nos pontos de acesso web (url) de banco de dados de triplas. Esse ponto de acesso é denominado de Sparql EndPoints. Endpoints genéricos executam consultas em qualquer dado RDF com acesso pela Web bastando declarar no FROM a localização do grafo RDF que contém o conjunto de triplas a ser inspecionado. Endpoints específicos executam consultas em conjuntos de dados particulares, fixados pela aplicação. Um SPARQL endpoint executa as consultas e retorna os resultados com um conjuntos de triplas. Esse resultado é um conjunto de tuplas (como linhas de uma tabela), que podem apresentar em diferentes formatos como, por exemplo, HTML, XML, JSON, CSV.. (WEB, 2016).

Como exemplo de EndPoints pode-se citar o projeto Linked Open Data (LOD) que disponibiliza um conjunto de repositórios RDF interligados, acessíveis por diversos Sparql Endpoints. Uma lista de EndPoints (http://pt.dbpedia.org/sparql; http://dbpedia.org/sparql; http://de.dbpedia.org/sparql; http://data.gov.uk/sparql...) pode ser consultado em https://www.w3.
.

3.1.7 Linked Data

Como segmento do desenvolvimento da Web Semântica, o Linked Data (LD) ou dados conectados é um conjunto de boas práticas para publicar e conectar dados estruturados na Web. Estas estão sendo utilizadas levando à criação do que conhecemos como Web of Data(WD). Possibilitando conexões de dados, em abrangência global, de distintas áreas, como de pessoas, companhias, livros, publicações científicas, filmes, músicas e tantos outros domínios na Web. Tecnicamente, Linked Data diz respeito aos dados disponíveis na Internet que são compreendidos por máquinas assim como pelas pessoas, com significado definido, ligado a outros conjunto de dados externos e que por sua vez, está conectado a outro conjunto de dados (BIZER TOM HEATH, 2009).

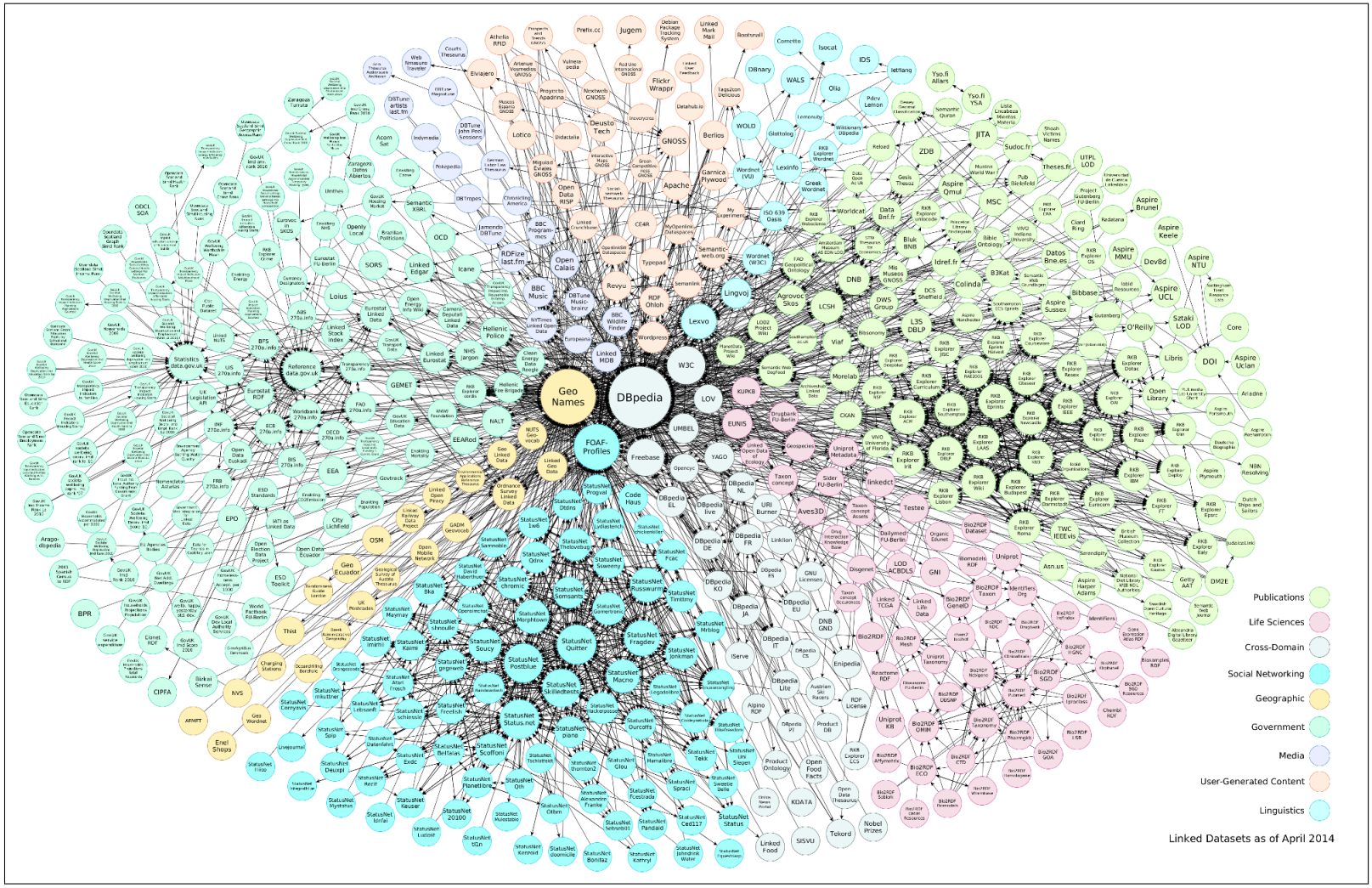
De acordo com (BERNERS-LEE; HENDLER; LASSILA, 2001), o conjunto das melhores hábitos a respeito do LD são:

- Usar URI como nome para recursos;
- Usar URI's HTTP para que as pessoas possam encontrar esses nomes;
- Quando alguém procura por uma URI, garantir que informações úteis possam ser obtidas por meio dessas URI, as quais deve estar representadas no formato RDF;
- Incluir links para outros URIs de forma que outros recursos possam ser conhecido;

Como exemplo da aplicação dos princípios dos dados ligados (LD), temos o projeto Linking Open Data (LOD), que tem como objetivo identificar conjuntos de dados disponíveis sob licenças abertas e convertê-los para RDF de acordo com os princípios Linked Data (BIZER TOM HEATH, 2009). A seguir é exibido esboço, de abril de 2014, do alcance e escala da Web of Data proveniente do projeto LOD, no qual cada nó no diagrama representa um conjunto de dados distinto publicado a partir dos padrões de Linked Data.

A figura 23 mostra os conjuntos de dados que são publicados como dados ligados à Web, bem como as relações de ligação entre os conjuntos de dados. O diagrama contém 570 conjuntos de dados e 2.909 relações de ligação entre os conjuntos de dados(PLANETDATA, 2014).

Figura 23 – Linked Datasets as of April 2014



Fonte: (PLANETDATA, 2014)

Dentre esses dados que participam do projeto LOD, podemos destacar a DBpedia. A DBpedia é um projeto que busca extrair e disponibilizar na Web informações estruturadas da enciclopédia aberta Wikipedia com o uso de tecnologias de Web Semântica e Linked Data. A versão em Inglês da base de conhecimento DBpedia descreve 4,58 milhão de coisas, dos quais 4,22 milhões são classificados em uma ontologia, incluindo 1.445.000 pessoas, 735.000 lugares (incluindo 478.000 lugares povoadas), 11.000 obras (incluindo 123.000 álbuns de música, 87.000 filmes e 19.000 jogos de vídeo), 241.000 organizações (incluindo 58.000 empresas e 49.000 estabelecimentos de ensino), 251.000 espécies e 6.000 doenças. Além disso, fornece versões localizadas do DBpedia em 125 idiomas, incluindo a portuguêsa. DBpedia está conectado com outros conjuntos de dados referenciados por cerca de 50 milhões de ligações RDF. Cada entidade na DBpedia possui uma URI com a sua descrição RDF, esta URI tem como base o endereço da Wikipedia (as informações disponíveis sobre o país Brasil na página da Wikipedia "http://pt.wikipedia.org/wiki/Brasil" está representada no DBpedia na URI "http://pt.dbpedia.org/page/Brasil"). A DBpedia atualmente é o principal hub (concentrador) de interligação da Web of Linked Data (Web de dados).

3.2 Anotação Semântica

Atualmente, a Web Tradicional (WT) é formada por páginas denominas de HTML e a estrutura da Web Semântica (WS) formada por OWL ou RDF. Com o objetivo de criar um documento na WT que seja passível de interpretação humana, seja analisado e processado por máquinas e softwares de modo a realizarem pesquisas precisas surge proposta da Anotação Semântica (AS) (FONTES; MOURA; CAVALCANTI, 2010).

Web services é considerado em geral, uma boa solução para o desenvolvimento de aplicações complexas distribuídas, tais como e-commerce e comunicações. Mas, devido a questão de que esses serviços são descritos pelos padões WSDL, os quais geralmente são sintáticos, faltando informações semânticas que proporcionaria respostas exatas, a AS é essencial para fornecer os componentes que faltam para o Web service (ZHANG; CHEN; FENG, 2013).

Anotação semântica é uma abordagem para alcançar os conceitos da Web semântica, cuja organização de informações fornece um meio, por onde a conexão lógica dos termos estabelece interoperabilidade entre sistemas. Ela é um esquema para geração e uso de metadados, habilitando novos métodos de acesso a informação. Uma anotação semântica é uma associação entre as expressões ou termos relevantes de um documento e os conceitos descritos em uma ontologia (BELLOZE et al., 2012).

Segundo (ELLER, 2008), uma anotação semântica de um documento descreve o seu conteúdo pela associação de palavras relevantes do texto e conceitos presentes na

ontologia. O resultado de uma anotação **A** é uma tupla (**as, ap, ao, ac**),onde: **as** é o dado anotado; **ao** é a anotação em si; **ap** é o predicado que define o tipo de relacionamento entre o **as** e **ao**; **ac** é o contexto em que a anotação foi feita (OREN et al., 2006).

Para que um documento Web seja bem anotado, é primordial a utilização de múltiplas ontologias ou taxonomias, por isso é essencial uma análise prévia da compatibilidade das ontologias com os domínios dos documentos.

Linguagens para anotação semântica em documentos da internet estão sendo utilizadas, são propostas como as linguagens RDFa, Microformatos e Microdata. Todas se caracterizam por utilizar um conjunto de atributos oriundos de um vocabulário, marcando trechos de um documento HTML ou xHTML, através de triplas semelhantes às utilizadas em RDF (FONTES et al., 2010). Nessa questão, destaca-se o RDFa por oferecer recursos complexos com *Blank nodes* e o uso de vocabulários arbitrários (FONTES; MOURA; CAVALCANTI, 2010). A vantagem do RDFa é que ele é recomendado pela W3C e disponibiliza suporte para *Blank nodes* (FONTES; CAVALCANTI; MOURA, 2013). RDFa possui os benefícios do RDF, possuindo recomendação pela W3C para interoperabilidade e legibilidade dos dados, é considerado flexível e semanticamente melhor que os microformatos. Além disso, marcações em RDFa permitem estender funcionalidades, tal como links de URL para as entidades anotadas. Por isso RDFa é melhor do que Microformatos para anotação de meta-dados (VIRGILIO et al., 2013) (FONTES et al., 2010).

As figuras 24 e a tabela 1 representam um exemplo da aplicação da linguagem RDFa: nos marcadores HTML foram adicionados atributos, *typeof* que indica a Classe (Student) do sujeito (Celso) da relação e *property* que indica o predicado ou propriedade *hasName* e a string *Celso Fontes* que representa o objeto. E finalmente, o URI *http://ime.eb.br/vocabulary/* que representa o vocabulário que descreve o contexto em que as descrições do recursos estão definidos (VIRGILIO et al., 2013) e (FONTES et al., 2010). Os novos atributos não interferem na interpretação das informações pelas pessoas e permitem que as informações nos documentos sejam interpretados pelas máquinas.

Figura 24 – Exemplo RDFa.

```
<span typeof='ime:Student' about='#Celso'
xmlns:ime='http://ime.eb.br/vocabulary/'>
  Hi! My name is
    <span property='ime:hasName'>
        Celso Fontes</span>
</span>
```

Fonte: (FONTES; CAVALCANTI; MOURA, 2013)

Com relação às características e classificação da anotação semântica, tem-se que ela pode ser do tipo manual em que o usuário faz o processo de marcação do documento, selecionando as partes a serem anotada e descrevendo a anotação associada a um termo de

Tabela 1 – Mapa de atributos do RDFa.

Sujeito	**Predicado**	**Objeto**
about	property	content
src	rel	href
	ver	resource
	typeof	datatype

Fonte: Próprio Autor.

uma ontologia. O problema da anotação manual é a quantidade de erros, devido a falta de conhecimento ou familiaridade da pessoa que executa a anotação com o domínio, grau de formação, motivação pessoal e complexidade dos esquemas. E ainda, é um processo custoso e que não considera as várias perspectivas de uma fonte de dados (REEVE; HAN, 2005). Segundo (NETO, 2009), a criação de forma manual possui vários problemas: dificuldade na expressão do conhecimento, elevado consumo de tempo e passível de erro.

A outra características e classificação da anotação semântica é a do tipo automática, em que uma ferramenta executa a anotação sem a intervenção do usuário por meio do uso de técnicas como de processamento de linguagem natural (NLP), aprendizado de máquina, extração de informações entre outros, para associar as marcações e as expressões da ontologia. A anotação automatizada oferece a escalabilidade necessária para fazer anotações em documentos existentes na Web, e reduz a carga de anotar novos documentos. Outro benefício é a utilização de múltiplas ontologias para anotar um documento (REEVE; HAN, 2005). Existem, anotações com suporte manual e automática que são denominadas de híbridas. Outra característica importante é como as anotações são salvas: podendo ser de forma intrusiva (quando as marcações ou anotações são armazenadas no documento) e de forma não intrusiva (se as anotações são armazenadas em outro arquivo), não modificando o documento original. De acordo com (NETO, 2009), a criação de forma automática possui problemas como: remoção de ambiguidade, obtenção formal de descrições satisfatórias para os conceitos, entre outros.

Na geração semiautomática, é obrigatória a intervenção do usuário (especialista em gestão do conhecimento) em alguma parte da criação da ontologia ou da anotação.

A tabela 25 representa um resumo das ferramentas de AS com suas características.

Figura 25 – Quadro resumo das ferramentas de Anotação Semântica.

Characteristics of tools. Kind of annotation (A=automatic, H=hybrid, M=manual), Saved annotation (I=intrusive, NI=non-intrusive), Platform (D=desktop, W=web).

Tool	Kind of annotation	Saved annotation	Format of input documents	Format of ontologies	Arbitrary ontology	Platform
Annotea	M	NI	Web documents	-	No	W
Annozilla	M	NI	Web documents	-	No	W
Autômeta	H	I	TXT	N-Triple, RDF, OWL, XML	Yes	D
GATE	H	NI	PDF, TXT, HTML, DOC, ODT	RDF, OWL	Yes	D
GoNTogle	H	NI	PDF, RTF, TXT, DOC, ODT	OWL	Yes	D
KIM	A	NI	HTML	RDF, OWL	No	W
Knowtator	M	NI	PDF, TXT, HTML, DOC, ODT	RDF, OWL, XML	Yes	D
Melita	M	NI	PDF, TXT, HTML, DOC, ODT	OWL	No	D
MnM	M	NI	HTML, TXT	DAML + OIL, RDF	Yes	W
Ontea	A	NI	PDF, TXT, DOC, e-mails, e-mail attachments in HTML	OWL	No	D
RDFaCE	M	I	PDF, TXT, HTML, DOC, ODT	-	No	D
RDFa Editor	A	NI	PDF, TXT, HTML, DOC, ODT	RDF, OWL, XML	Yes	D
Yawas	M	I	Web pages	-	No	W

Fonte: (BELLOZE et al., 2012)

De acordo com (BELLOZE et al., 2012) explica-se de forma detalhada as características de cada uma dessas ferramentas:

1. Annotea: é um projeto da W3C. As anotações desta ferramenta referem-se a comentários, anotações, explicações ou comentários gerais de documentos Web. Ele é parte dos esforços da Web semântica e usa um esquema de anotação baseado em RDF. Os metadados das anotações são armazenados localmente ou em servidores de anotação.

2. Annozilla: é semelhante ao Annotea, porém funciona como um plugin do browser Mozilla Firefox e armazena as anotações em RDF em um servidor. As anotações são destacadas para o usuário mesmo quando a página é recarregada.

3. AutôMeta (Automatic Metadata annotation tool): permite a anotação de um ou mais documentos usando uma ontologia previamente selecionada. As anotações geradas pela ferramenta são armazenadas usando o padrão RDFa.

4. GATE (General Architecture for Text Engineering): é uma ferramenta para aplicações de processamento de linguagem natural. Ele integra um ambiente de desenvolvimento que inclui plugins e outros componentes que permitem tanto a anotação quanto a extração de informações.

5. Gontogle: é uma ferramenta para anotação e pesquisa. Ele fornece maneiras de pesquisar usando uma combinação de busca semântica e as palavras-chave. As anotações são salvas como uma instância no servidor de ontologia e adicionados a uma lista do editor de anotações.

6. KIM: é baseado em uma plataforma Web para pesquisa semântica, anotações de dados e documentos. Ele possui a sua própria ontologia que inclui entidades de interesse geral. O acesso aos recursos da plataforma KIM é realizado através de uma interface Web (KIM Web), que permite métodos tradicionais de pesquisa por palavra-chave ou busca semântica (entidades, padrões).

7. Knowtator: é um plugin do Protégé, e permite um incremento de ontologias para adaptar a aplicação do usuário. A anotação é feita sobre a região do texto selecionado e a especificação da ontologia utilizando as presentes no Protégé.

8. Melita: é uma ferramenta que tem a sua própria ontologia, permitindo aos usuários adicionar os seus resultados para a ontologia, aumentando-a em cada anotação satisfatória.

9. MnM: é uma ferramenta que permite anotações em páginas da Web. Ela utiliza um algoritmo de aprendizado sobre as anotações para posteriormente calcular a precisão e novamente chamar as anotações no corpus. Ele integra um navegador da Web com um editor de ontologia e fornece APIs (Interface de Programação de Aplicativo) para conexão entre servidores de ontologias e ferramentas de extração de informação.

10. ONTEA: utiliza as suas próprias ontologias que estão relacionadas somente a endereços, nomes e e-mails.

11. RDFaCE (RDFa Content Editor): é um plug-in para TinyMCE Javascript Editor WYSIWYG que permite a anotação intrusiva no padrão RDFa. Em vez de ontologias, usa APIs que sugerem os recursos para a anotação. Esses recursos fornecem as URIs para objetos, propriedades e namespaces.

12. RDFa Editor: apresenta-se como uma ferramenta promissora que usa o padrão RDFa para as anotações. Ela permite a utilização arbitrária de ontologias.

13. Yawas: é um plugin desenvolvido para os browsers Firefox e Google Chrome, onde as anotações são destacadas nas páginas web, mas sem usar recursos semânticos.

No trabalho do (REEVE; HAN, 2005), aborda-se a classificação das Plataforma de Anotação Semântica (PAS), com uma revisão da arquitetura dessas plataformas, suas abordagens e performance. Como demonstrado na figura 26, ele classifica a Plataforma de Anotação Semântica (PAS) em duas categorias: A primeira, **baseada em padrões**: deve ser provido para a plataforma um conjunto inicial de entidades, para que no processo de leitura do corpus sejam encontrados padrões existentes nas entidades. Novas entidades são descobertas, juntamente com os novos padrões. O processo é recursivo até que não haja mais entidades a serem encontradas ou o processamento seja interrompido pelo usuário. Nessa classificação também estão anotações criadas a partir de regras geradas manualmente para encontrar entidades em um texto. E a segunda, **baseado em aprendizagem de máquina**: esse possui dois métodos, o probabilístico que utiliza modelos estatísticos para prever e identificar as entidades do texto e o indutivo que reutiliza um processo de extração de informações para induzir a identificação de entidades. A arquitetura classifica-se em **não extensiva** que se concentra em um único domínio, método ou kit de ferramenta; enquanto a **extensiva** permite que os componentes do sistema possam ser substituídos ou estendidos com outros componentes (isso permite que novos métodos de anotação sejam testados e integrados enquanto reutiliza todos os outros recursos da plataforma).

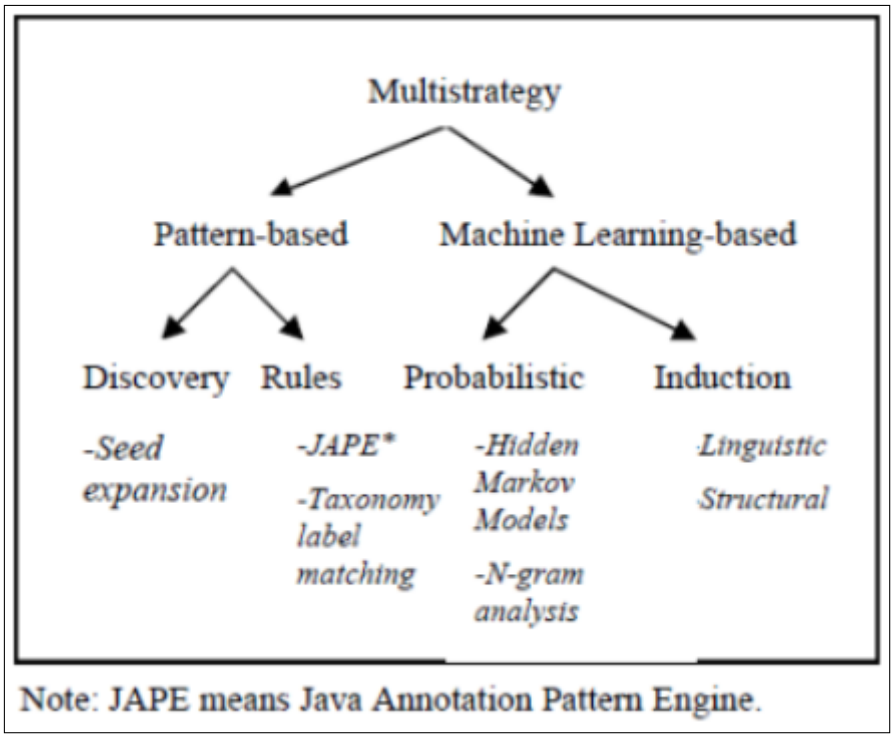
Nesse seu trabalho é descrito brevemente, com a utilização de uma revisão de literatura, PAS afim de mostrar suas arquiteturas e desempenhos medidos empiricamente:

AeroDAML: Possui uma abordagem baseada em padrões, mapeia nomes próprios e as relações comuns com as classes e propriedades correspondentes na ontologia. AeroDAML executa o AeroText, uma API Java, utilizada para acessar o extrator de informações (IE) e mapeá-los em RDF triplica usando uma ontologia como guia.

Armadillo: adota uma abordagem baseada em padrões para encontrar entidades. Ele utiliza o sistema Amilcare como extrator de informações para realizar a identificação de entidades das páginas web de modo indutivo. Uma vez que as origens são encontradas, o padrão de entidades é utilizado para descobrir entidades adicionais.

KIM (Knowledge and Information Management): Contém uma ontologia, uma base de conhecimento, uma anotação semântica, um servidor de indexação e recuperação, bem como front-end para interface com o servidor. O processo de anotação semântica depende de uma ontologia denominada KIMO e de uma base de conhecimento entre domínios. Com o KIMO define-se um conjunto de classes de entidades, relaciona-

Figura 26 – Classificação das Plataformas de Anotação Semântica



Fonte: (REEVE; HAN, 2005)

mentos e restrições de atributos. Entidades encontradas durante o processo de anotação são compatíveis com seu tipo na ontologia e com a base de conhecimento, esse mapeamento duplo permite que o processo de extração de informações seja melhor, fornecendo dicas de desambiguação com base em atributos e relações. A extração de informação é realizada utilizando componentes do kit de ferramentas do GATE.

MnM: disponibiliza um ambiente para anotar manualmente um corpus como exemplo para depois alimentar o sistema de indução que tem como base um algoritmo de processamento de linguagem natural. É uma biblioteca de regras induzidas que pode ser usado para extrair as informações de corpus de textos.

MUSE: foi implementado utilizando o framework GATE e possui a capacidade de realizar reconhecimento e conferencia do nome de entidades. Possui um recurso de processamento para extração de informações (IE) que permite obter precisão semelhante aos sistemas de aprendizado de máquina. Esse sistema é mais sofisticado do que um dicionário de palavras, porque essa não pode fornecer uma lista exaustiva de todas as potenciais-entidades nomeadas, e não pode resolver as entidades ambíguas.

Ont-O-Mat: é um framework de anotação semântica. Seu extrator de informações, Amilcare, é baseado em máquina de aprendizado que exige uma formação corpus de documentos anotados manualmente. Ele utiliza o ANNIE ("ANearly-New IE system") que é parte do integrante do GATE. O resultado do processamento do Annie é passado

para Amilcare, que induz as regras para a IE usando um algoritmo.

SemTag: é uma plataforma abrangente para realizar uma escala de anotação de páginas da web. Suas anotações são geradas e armazenadas separadamente do documento de origem. A inteção dessa plataforma é fornecer um repositório público com uma API que permitirá que agentes recuperem a página web de sua fonte e, em seguida, solicitem as anotações separadamente.

Na figura 27, (REEVE; HAN, 2005) demonstra os atributos que tem impacto sobre a anotação semântica automatizada. O método utilizado para localizar entidades é o principal determinante no desempenho. Os métodos de aprendizado de máquina, tais como aqueles usados por Amilcare, geralmente possuem um desempenho melhor, embora o sistema MUSE que é baseado em regras, utilizando processamento condicional, mostrou que esse sistema baseado em regras pode igualar o desempenho com sistema baseado em aprendizagem de máquina. Sistemas baseados em regras exigem regras, os sistemas de descoberta padrão requerem um conjunto inicial de dados, sistemas de máquinas de aprendizagem requerem um corpus de treinamento (normalmente anotado), enquanto outros exigem a construção de dicionários para o reconhecimento da entidade nomeada.

Figura 27 – Resumo das Características das Plataformas de Anotação Semântica

Platform	Method	Machine Learning	Manual Rules	Bootstrap Ontology
AeroDAML [14]	Rule	N	Y	WordNet
Armadillo [10]	Pattern Discovery	N	Y	User
KIM [18]	Rule	N	Y	KIMO
MnM [21]	Wrapper Induction	Y	N	KMi
MUSE [16]	Rule	N	Y	User
Ont-O-Mat: Amilcare [12]	Wrapper Induction	Y	N	User
Ont-O-Mat: PANKOW [5]	Pattern Discovery	N	N	User
SemTag [9]	Rule	N	N	TAP

Fonte: (REEVE; HAN, 2005)

A maioria das PAS lidam com um sistema de extração de informação (IE) ex-

terna, das quais a maioria foi desenvolvida a partir da comunidade de processamento de linguagem natural (PLN). Alguns sistemas de IE dispõe de serviços adicionais, tais como reconhecimento de entidades nomeadas (NER), IE por regras de indução usando a máquina de aprendizagem e encontram relações de identidade entre as entidades em texto (co-referenciação).

Com relação a performance dessas ferramentas, nas suas observações e de acordo com a figura 28, é observado que o desempenho melhor foi realizado pela PAS MnM que é baseada em aprendizagem de máquina. A PAS baseada em padrões com o pior desempenho foi o MUSE. E a pior de todas as PAS foi o Ont-O-Mat. A medida padrão de *Precision, Recall*, e *F-measure*, foram obtidas a partir do campo recuperação da informação, utilizados pelos autores das PAS na determinação da eficácia de anotação.

Figura 28 – Performance das Plataformas de Anotação Semântica

Framework	Precision	Recall	F-Measure
Armadillo	91.0	74.0	87.0
KIM	86.0	82.0	84.0
MnM	95.0	90.0	n/a
MUSE	93.5	92.3	92.9
Ont-O-Mat: PANKOW	65.0	28.2	24.9
SemTag	82.0	n/a	n/a

Fonte: (REEVE; HAN, 2005)

(REEVE; HAN, 2005) conclui a sua pesquisa afirmando que as Plataformas de Anotação Semântica (PAS) podem ser distinguidas pelo seu método de anotação, sendo esse o componente que tem a maior relevância sobre a eficácia de uma anotação semântica. Os algoritmos, de aprendizado de máquina, possuem um desempenho eficaz em relação aos métodos baseados em padrões, mas foi demonstrado na pesquisa que um sistema baseado em regras utilizando o processamento condicional pode executar tão bem quanto um sistema de aprendizagem de máquina. A contínua evolução das PAS e a ampliação de novos recursos afim de proporcionar uma anotação ótima é fundamental para a realização da Web Semântica.

De acordo com (DERCZYNSKI DIANA MAYNARD, 2014) o reconhecimento de entidades mencionadas (NER) é uma tarefa crítica na extração de informação (IE), uma vez que identifica quais trechos do texto são menções de entidades no mundo real. Ele é

um pré-requisito para a tarefa de IE. NER é difícil em conteúdo gerado pelo usuário, e no gênero microblog especificamente, por causa da quantidade reduzida de informações contextuais em mensagens curtas. Esse autor apresenta em seu trabalho a figura 29 no qual elucida as características de algumas das ferramentas de extração de entidades descritas e ou mencionadas no seu trabalho. Para cada sistema ele indica que tipo de abordagem é usada, quais são os idiomas suportados, domínio geral ou específico, o número de classes de tipos de entidades são classificados, como o sistema pode ser usado (para download ou, por exemplo, através de um serviço web), que licença se aplica ao sistema e se o sistema pode ser adaptado. Observa-se que para AlchemyAPI, Lupedia, Saplo, Textrazor e Zemanta não são sistemas gratuitos possuindo algoritmos e os recursos restritos.

Figura 29 – Ferramentas de Reconhecimento de Entidade

Feature	ANNIE	Stanford NER	Ritter et al.	Alchemy API	Lupedia
Approach	Gazetteers and Finite State Machines	CRF	CRF	Machine Learning	Gazetteers and rules
Languages	EN, FR, DE, RU, CN, RO, HI	EN	EN	EN, FR, DE, IT, PT, RU, ES, SV	EN, FR, IT
Domain	newswire	newswire	Twitter	Unspecified	Unspecified
# Classes	7	4, 3 or 7	3 or 10	324	319
Taxonomy	(adapted) MUC	CoNLL, ACE	CoNLL, ACE	Alchemy	DBpedia
Type	Java (GATE module)	Java	Python	Web Service	Web Service
License	GPLv3	GPLv2	GPLv3	Non-Commercial	Unknown
Adaptable	Yes	Yes	partially	No	No
	DBpedia Spotlight	TextRazor	Zemanta	YODIE	NERD-ML
Approach	Gazetteers and Similarity Metrics	Machine Learning	Machine Learning	Similarity Metrics	SMO and K-NN and Naive Bayes
Languages	EN	EN, NL, FR, DE, IT, PL, PT, RU, ES, SV	EN	EN	EN
Domain	Unspecified	Unspecified	Unspecified	Twitter	Twitter
# Classes	320	1779	81	1779	4
Taxonomy	DBpedia, Freebase, Schema.org	DBpedia, Freebase	Freebase	DBpedia	NERD
Type	Web Service	Web Service	Web Service	Java (GATE Module)	Java, Python, Perl, bash
License	Apache License 2.0	Non-Commercial	Non-Commercial		GPLv3
Adaptable	Yes	No	No	Yes	Partially

Fonte: (DERCZYNSKI DIANA MAYNARD, 2014)

O trabalho de (DERCZYNSKI DIANA MAYNARD, 2014) realiza um comparativo entre os sistemas NER apresentados na figura 29. Observando essa imagem, o autor traz a luz duas opções de ferramenta (TextRazor e AlchemyApi) que auxiliam na tarefa de solucionar o problema dessa dissertação, pois possuem suporte para língua portuguêsa.

3.3 Currículo Lattes

De acordo com o CNPq, o Lattes é a base de dados de currículos, instituições e grupos de pesquisa das áreas de Ciência e Tecnologia. Desde os anos 80, havia interesse por parte dessa instituição de criar um formulário padrão para registrar os currículos dos pesquisadores brasileiros. A partir desse interesse, o sistema originou-se com a denominação

de Banco de Currículos que contava com a captação dos dados em papel e em seguida com a digitação no sistema. No início dos anos 90, passa a se chamar BCURR, onde a captação dos dados passa a ser em um formulário eletrônico dentro do sistema operacional DOS e depois era enviado via disquete para ser importado na base de dados. Tempos depois, esse sistema evolui sendo nomeado de Cadastro Nacional de Competências em Ciência e Tecnologia (CNCT) e caracterizado por possuir um formulário eletrônico no ambiente Windows e em seguida os dados eram enviados de forma off line por meio da Internet. No final dessa mesma década, os grupos (CESAR - Centro de Estudos e Sistemas Avançados do Recife - da Universidade Federal de Pernambuco, e o grupo Stela - atual Instituto Stela - da Universidade Federal de Santa Catarina) desenvolvem uma única versão com a capacidade de integrar as existentes. Em meados do ano de 1999, o CNPq padronizou e lançou o Currículo Lattes para ser o formulário de currículo da esfera do Ministério da Ciência e Tecnologia e CNPq.

A partir de 2001, começou a discussão com relação à abertura e padronização XML com relação ao Lattes. Algumas universidades como UFSC, UNICAMP, UFRJ, USP, UFRGS, UFBA e UFRN solicitaram ao CNPq a abertura tecnológica das informações dessa plataforma. A partir disso, originou a construção da Linguagem de Marcação da Plataforma Lattes (LMPL), sob coordenação da CGINF/CNPQ, sendo os trabalhos de desenvolvimento conduzidos pelo Grupo Stela da UFSC. Mais adiante, esse trabalho resultou na formação da Comunidade Virtual LMPL, que definiu o modelo DTD (Data Type Definition) XML do Currículo Lattes. Esse padrão XML foi desenvolvido inicialmente utilizando a linguagem de definição de tipos, DTD (Document Type Definition). Em seguida, com a homologação da linguagem XML Schema pelo Consórcio W3C, a comunidade CONSCIENTIAS-LMPL construiu uma nova estrutura utilizando a linguagem de esquemas para o mesmo padrão XML de Currículo Vitae.

Com isso, tornou-se viável a partir da versão 1.4 do Lattes a abertura da Plataforma, do ponto de vista de conteúdo dos dados, ficando inalterado o acesso técnico às informações, preservando a segurança dos pesquisadores.

O Lattes é um padrão nacional da vida pregressa e atual dos pesquisadores e estudantes. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, tornou-se um elemento essencial à análise de mérito e competência das solicitações de financiamentos na área de ciência e tecnologia. É um sistema estratégico para as atividades de planejamento e gestão. É utilizado na formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação.

Qualquer pessoa pode preencher o seu currículo, acessando o site da Plataforma Lattes. Os dados preenchidos são armazenados e disponibilizados publicamente na Internet, tanto em formato HTML quanto em XML. O currículo armazenado na base Lattes recebe um número identificador.

A Plataforma disponibiliza as seguintes funcionalidades:

1. Busca de Currículos: localiza currículos utilizando diversos filtros (como nome, titulação, palavras-chaves etc).
2. Rede de Colaboração: É exibido um grafo no qual os vértices são os pesquisadores e as arestas as colaborações com outros pesquisadores.
3. Painel Lattes: Contém dados estatísticos da base da Plataforma, disposto em forma de gráficos.

Mas, a Plataforma carece de funcionalidades que explorem os dados de um grupo de currículos. Certas informações são difíceis de se obter, como por exemplo:

1. Quais professores/pesquisadores publicaram em um determinado ano? Destes, quantas publicações em conferências internacionais? E nacionais?

2. Quais são os professores/pesquisadores de um Departamento ou Universidade? Quais destes possuem registro de publicação? Quais publicaram livros? Quais publicaram capítulos de livros?

3. Quais são os professores/pesquisadores que publicaram em coautoria com um pesquisador/professor?

4. Quais são as teses de doutorado e dissertações de mestrado finalizadas sob orientação de algum professor do grupo nos últimos X anos?

5. Se comparado às publicações com anos anteriores, está havendo um decréscimo ou crescimento no número de publicações de uma determinada Universidade/Faculdade?

6. Os dados de orientações informados por um pesquisador/professor estão condizentes com os informados pelos orientados?

Trabalhos mencionados na próxima seção foram criados com o objetivo de disponibilizar ou facilitar o uso dessas informações.

3.3.1 Plataforma Lattes e a Web Semântica

Durante o desenvolvimento deste trabalho uma pergunta se fez necessária: "Quais são os trabalhos que associa à Web Semântica com a Plataforma Lattes (PL)?". Diante dessa questão temos:

O trabalho desenvolvido em 2002 por (BONIFACIO, 2002), *Ontologias e consulta semântica: uma aplicação ao caso Lattes*, ele introduz conceitos básicos, uma introdução a novos paradigmas de linguagens e ferramentas que estão dando os primeiros passos em direção a Web Semântica na PL. Um processo de tradução semi-automática dos dados do documento XML gerado pela exportação do Sistema Lattes para o modelo ontológico

em DAML+OIL foi apresentado, sendo que a contribuição desenvolvida nesse trabalho foi permitir uma melhor compreensão dos conceitos, linguagens e ferramentas que foram apresentadas, com a aplicação deles no caso do Currículo Lattes. Como resultado foi elaborado uma proposta de uma ontologia para a plataforma na linguagem DAML+OIL, denominada de OntoLattes.

O Projeto de (CASTAñO, 2008), *Populando ontologias através de informações em HTML - o caso do currículo lattes*, foi um trabalho de dissertação de mestrado que foi utilizada como fonte de informações os currículos da PL para popular automaticamente uma ontologia (criada por Ailton Sergio Bonifacio e depois convertida de DAML+OIL para OWL por Marcos Yoshinori Nakashima) e utilizá-la como uma base de dados a ser consultada para geração de relatórios. Todo processo de extração de informações (*wrappers*) foi executado a partir de documentos HTML, com processamento posterior para inserção correta na ontologia, de acordo com sua semântica. Nesse processo foi encontrado dificuldades ou problemas como: identificar corretamente os textos dos arquivos originais para que fosse possível mapear a ontologia com a semântica correta dos termos, identificar e retirar as duplicidades de instâncias que se referem a um mesmo objeto. No trabalho foi utilizado duas abordagens na busca por similaridades e demonstrado suas características principais. Também foi exemplificado de forma superficial uma comparação da criação de consultas em SPARQL, XQuery e SQL.

O trabalho de (GALEGO, 2013), *Extração e Consulta de Informações do Currículo Lattes Baseada em Ontologias*, foi apresentado em uma revisão de trabalhos que propuseram a geração de relatórios sumarizados de um grupo de pessoas, alguns com desenvolvimento de ontologias, no domínio do Lattes. Ele descreve sobre o OntoLattes, que foi a construção de uma ontologia, no formato OWL, para comportar os dados dos currículos dos pesquisadores; sobre o SemanticLattes que realiza as tarefas de importação de currículos e lista de veículos de publicações científicas em duas ontologias (descritas inicialmente em DAML+OIL e em seguida OWL), permitindo consultas às instâncias, ele possui um motor de busca que processa a pergunta em linguagem natural e o software, por meio de identificação das palavras-chave, reconhece a pergunta e faz a respectiva consulta em SPARQL. Sobre o ScriptLattes que é um software que cria relatórios gerenciais obtidos a partir de um conjunto de currículos em formato HTML ou XML, o ScriptLattes não trabalhou com ontologia. As estruturas de dados foram construídas utilizando o conceito de orientação à objetos. Ele foi de relevância no mundo acadêmico e científico, sendo confundindo muitas vezes com uma ferramenta que foi desenvolvida e disponibilizada pelo CNPq; e por fim, o projeto Sucupira, que tem por objetivo a extração de informações da Plataforma Lattes para identificação de redes sociais acadêmicas. Uma das principais funcionalidades deste sistema, Sucupira, é o gerenciamento de uma lista de pesquisadores definida pelo usuário, sendo possível visualizar um mapa contendo o endereço profissional dos pesquisadores, um gráfico sumarizado do número de publicações por ano e tipo, e um

grafo relacionando os pesquisadores a outros currículos.

Além dessa revisão, (GALEGO, 2013), desenvolve uma ferramenta denominada de Dynamic Lattes, que reutiliza as funcionalidades dos trabalhos citados anteriormente e incorpora outras funcionalidades como a possibilidade de alteração do conteúdo dos dados do relatório sem necessidade de alteração da apresentação, a inclusão do relatório de dados inconsistentes, possibilidade de associar uma orientação a formação de algum membro e resumo da comparação dos dados informados pelo orientador com o orientado.

Uma das questões a ser observada nesse levantamento é com relação ao núcleo das pesquisas mencionadas, conceitos de web semântica e ontologia, não foi encontrado pesquisas sobre a anotação semântica utilizando Linked Open Data. Também pode-se destacar a sugestão de futuro trabalho mencionado por esse último autor,(GALEGO, 2013) que é, "Explorar as funcionalidades de Linked Data para que seja possível integração com outras bases de conhecimentos.", que vem ao encontro com o núcleo desta pesquisa (anotação semântica com Linked Open Data).

3.4 Elaboração da RSL

O objetivo da Revisão Sistemática da Literatura foi selecionar, analisar e interpretar produções pertinentes para um problema de pesquisa. Um protocolo de avaliação foi criado para identificar a necessidade da revisão e os critérios de busca que serão utilizados.

Para essa dissertação, a necessidade da revisão permitiu o entendimento dos conceitos e como está a situação do tema dessa pesquisa, anotação semântica, no quesito anotação automática de documentos Web com a utilização do LOD. Fica definido nesta seção que a revisão da literatura tem como proposição responder as seguintes questões:

1. : Q1: Quais são os componentes/módulos que são fundamentais para montar uma arquitetura de anotação semântica com Linked Open Data?
2. : Q2: Com quais ferramentas/frameworks podemos efetuar o processo de anotação semântica e extração de entidades?
3. : Q3: Dos componentes/módulos encontrados, qual(is) permite executar a ação de anotar com eficiência?

De acordo com figura 30, Protocolo da Revisão Sistemática de Literatura, ficou definido o protocolo de revisão sistemática da literatura definida (PRSL).

Figura 30 – Protocolo da Revisão Sistemática de Literatura

Protocolo da revisão sistemática de literatura	
Bases de Pesquisas	
	Acm - & http://dl.acm.org/
	IEEE Xplore - http://ieeexplore.ieee.org
	ScienceDirect - http://www.sciencedirect.com/
	Springer - http://link.springer.com/
Critério de Pesquisa	annotation and Linked Data.
	annotation and Linked Data or Linked Open Data
	semantic annotation and Linked Data
Critério de Seleção	
	Periódicos revisado por pares (Journals, Magazines and Conference Publication)
	Grande área: Ciências exatas da terra.
	Período: 2010 - 2014.
	Inglês e Português.
	Texto Completo.
Critério de Exclusão	
	Não disponibilizar o texto completo para leitura.
	Não estar relacionado com anotação semântica.
	Livros, Resenhas, Monografias.
	Estudos duplicados.
	Análise das palavras chave no abstract e título.
	Análise das palavras chave no conteúdo.
Categorização Definida	
	Base
	Tipo Produção
	Ano
	País
	Contribuição
	Ferramenta Anotacao
	Ferramenta de Extracao de Informação
	Extensível
	Plataforma
	Método de Reconhecimento da Entidade
	Forma de Anotação de Entidades
	Modo Salva as Anotações
	Formato Salva Anotações
	Formato Entrada Doc1

Fonte: Próprio Autor.

Utilizando o critério de pesquisa e executando-o nas bases definidas foi possível efetuar o download automaticamente das produções selecionadas. Para o download automático foi utilizado o programa Zotero. Esse é uma extensão para o Firefox capaz de catalogar as páginas da internet automaticamente e organizar a coleção de itens selecionados. Esse programa permite armazenar sites com artigos, vídeos, arquivos em PDF e outras publicações disponíveis pela web de forma rápida e fácil. Na extração automatizada observou os critérios de pesquisa e seleção. Como o núcleo do trabalho envolve anotação automática de documentos a partir de dados abertos, ficou definido que as palavras-chave que serão utilizadas nas bases de pesquisas serão (annotation and Linked Data; annotation and Linked Data or Linked Open Data; semantic annotation and Linked Data). A idéia foi selecionar produções que abordem conceitos e práticas relacionadas a anotação semântica e Linked Open Data. As etapas de seleção das produções foi efetuada através da análise subjetiva, em seguida relevância das palavras chave no conteúdo, e por fim leitura das produções.

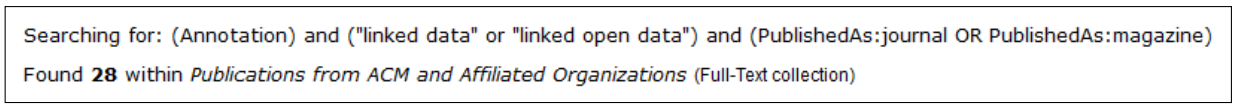
Nesta etapa de planejamento, o processo é iterativo e dinâmico enquanto a RSL estiver ocorrendo, os resultados são adaptados de acordo com os objetivos relacionados na revisão podendo eles serem delimitados de acordo com a evolução dos critérios.

3.4.1 Realização

Após o planejamento e definições mencionadas anteriormente, foi executado o processo de seleção e filtragem das publicações. Nessa etapa de realização da RSL é efetuada a seleção dos estudos primários de acordo com os critérios estabelecidos. Os critérios da pesquisa e seleção são ajustados para que se consiga alcançar um número de produções que seja relevante para o trabalho, abaixo tem-se as seguintes consultas e resultados para cada base.

Construção da pesquisa na Base ACM:

Figura 31 – String de pesquisa utilizada na base ACM



Fonte:Próprio Autor.

Searching for: (annotation) and ("linked open data"or "linked data") and (PublishedAs:journal OR PublishedAs:magazine) Sendo encontradas 28 produções.

Construção da pesquisa na IEEE:

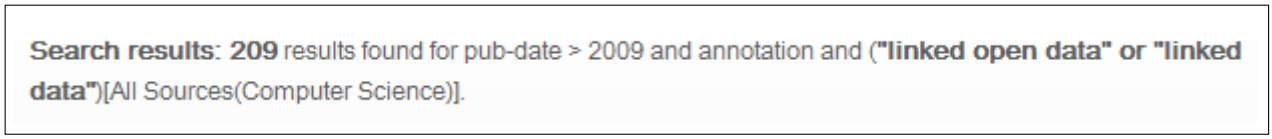
Figura 32 – String de pesquisa utilizada na base IEEE

Fonte: Próprio Autor.

You searched for: ((annotation) AND "linked open data"OR "linked data") You Refined by Content Type: Conference Publications Remove , Journals Magazines Remove Publication Year: 2010 - 2014. Tendo como resultado 322 produções.

Construção da pesquisa na ScienceDirect:

Figura 33 – String de pesquisa utilizada na base ScienceDirect



Fonte: Próprio Autor.

O resultado da pesquisa foi de: 209 produções encontradas para (annotation and ("linked data"or "linked open data")[All Sources(Computer Science)]).

Construção da pesquisa na Springer:

Figura 34 – String de pesquisa utilizada na base Springer

Fonte: Próprio Autor.

Foi localizado 240 produções com resultado para '(annotation and ("linked data"OR "linked open data"))' within Computer Science; Article; 2010 - 2015

Nessa etapa de seleção foi realizada uma análise das produções utilizando as palavras chave no abstract e título, utilizando um programa que foi criado na linguagem PHP com a finalidade de verificar se as palavras chaves existiam no abstract e título. O programa leu um arquivo «base».bib e retornou «base-excl01».bib. A tabela 2 representa o resultado dessa primeira seleção:

Tabela 2 – Filtro1: Lista de Produções Por Base de Pesquisa

Base	Selecionado	Excluído	Total
ACM	28	4	24
IEEE	321	3	318
SCIENCEDIRECT	229	137	192
SPRINGER	240	30	210

Fonte: Próprio Autor.

Em seguida foi utilizado a ferramenta AlchemyAPI para gerar dados com relação a importância das palavras chaves (annotation, semantic, linked Data, Linked Open Data) dentro do contexto. Foi criado um programa em PHP para utilizar a API do Alchemy: A entrada do programa foi um arquivo «base-excl01».bib (com os metadados das produções) e o seu retorno foi um arquivo «base-excl01Result».xls (com relevância das palavras chave para cada produção). Nesse, realizou-se uma análise subjetiva observando e selecionando as publicações que continham as palavras chaves em seu conteúdo e sempre observando a relevância desses termos dentro do documento.

Tabela 3 – Filtro2: Lista de Produções Por Base de Pesquisa

Base	Selecionado	Excluído	Total
ACM	24	21	5
IEEE	318	300	13
SCIENCEDIRECT	192	181	11
SPRINGER	210	171	39

Fonte: Próprio Autor.

Após efetuar a leitura completa dos artigos selecionados na tabela 3, foi selecionado para este trabalho 31 produções que estão descritas no apêndice A.

3.4.2 Resultados

Esta seção apresentará os resultados encontrados na RSL, sendo a base do trabalho dessa pesquisa condizente com o planejamento proposto. Foi executada a leitura e análise

dos artigos propostos na seção realização, e a seguir é apresentado uma análise do que foi identificado e pode vir a responder as questões levantadas.

Foi observado nas leituras das produções encontradas que o tema dessa pesquisa encontra-se evoluído no cenário internacional, mas no Brasil o tema está em discussão, existem poucas produções científicas e ferramentas nacionais desenvolvidas nesse assunto. Foram analisadas 01 da Índia, Egito, Reino Unido, Alemanha, Portugal, Romênia e Iran, 02 produções da China, Espanha e Sérvia, 05 Brasil, 03 da Grécia, Itália e 06 produções dos Estados Unidos. De maneira proporcional ferramentas de extração de entidades mencionadas nas pesquisas são em sua maioria destinadas as línguas inglesa, espanhola, italiana e alemã, comportando dessa mesma maneira os anotadores semânticos.

Na conjuntura brasileira o tema encontra-se bem recente com alguns trabalhos como (BELLOZE et al., 2012), (FONTES et al., 2010), (MUNARO; LIMA; CAMPOS, 2012), porém sem abrangência com o tema anotação semântica com as bases abertas (Linked Open Data).

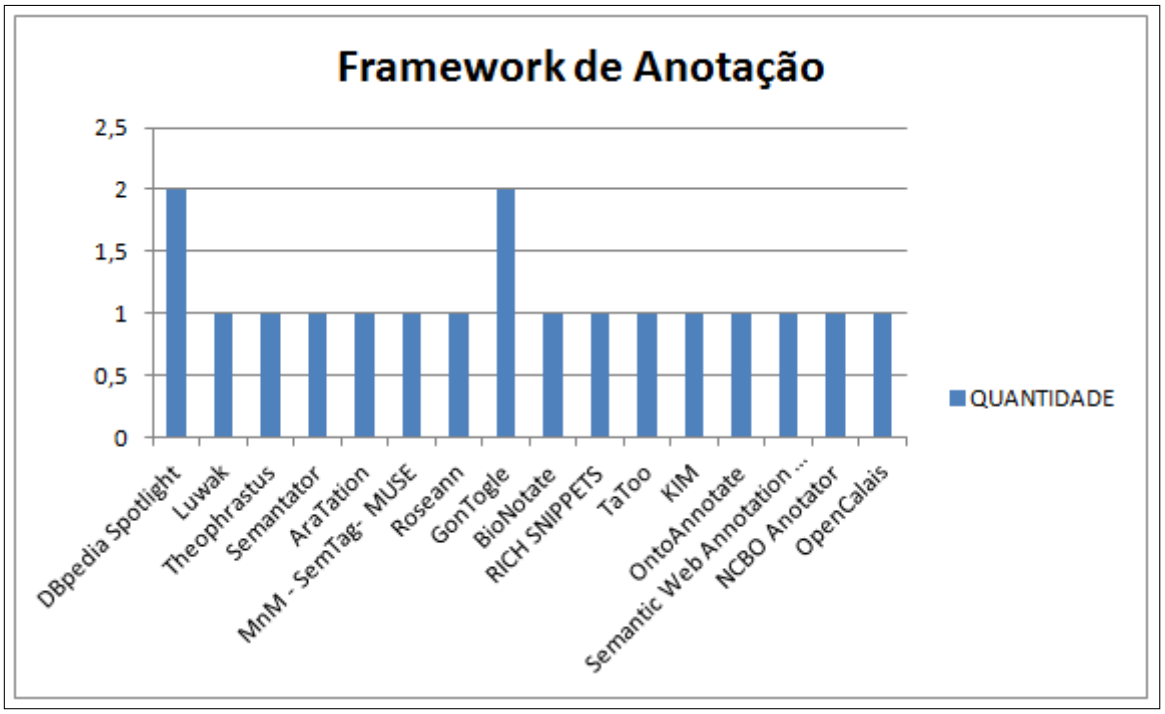
R1: Quais são os componentes/módulos que são fundamentais para montar uma arquitetura de anotação semântica com Linked Open Data?

Nas leituras realizadas, pode-se citar os autores (FAFALIOS; PAPADAKOS, 2014), (NETO, 2009), (FONTES; CAVALCANTI; MOURA, 2013) indicam um padrão quanto ao modelo de framework de anotação semântica: um objeto responsável pela análise e extração dos termos de um documento e um objeto responsável pela criação de um documento anotado. A variação encontra-se no quesito da base de dados utilizada para efetuar o mapeamento dos termos encontrados, que no caso (NETO, 2009), (FONTES; CAVALCANTI; MOURA, 2013) utilizam ontologias específicas para o desenvolvimento dos seus trabalhos e no caso do (FAFALIOS; PAPADAKOS, 2014) e (ZHANG; CHEN; FENG, 2013) utilizam as ontologias disponíveis no LOD. Quanto a forma de anotar e salvar as anotação foi encontrado a utilização do RDF e não intrusiva nas produções (TAO et al., 2013), (SALEH; AL-KHALIFA, 2009), (VIRGILIO et al., 2013) e RDFa e não intrusiva em (MENDES et al., 2011), também tem a produção (BUTUC, 2009) que utiliza RDF, JSON, Microformatos e não intrusiva.

R2: Com quais ferramentas/frameworks podemos efetuar o processo de anotação semântica e extração de entidades?

A figura 35 representa as ferramentas utilizadas nas pesquisas para anotar semanticamente os documentos ou em alguns casos foram mencionadas como solução para trabalhar com anotação semântica.

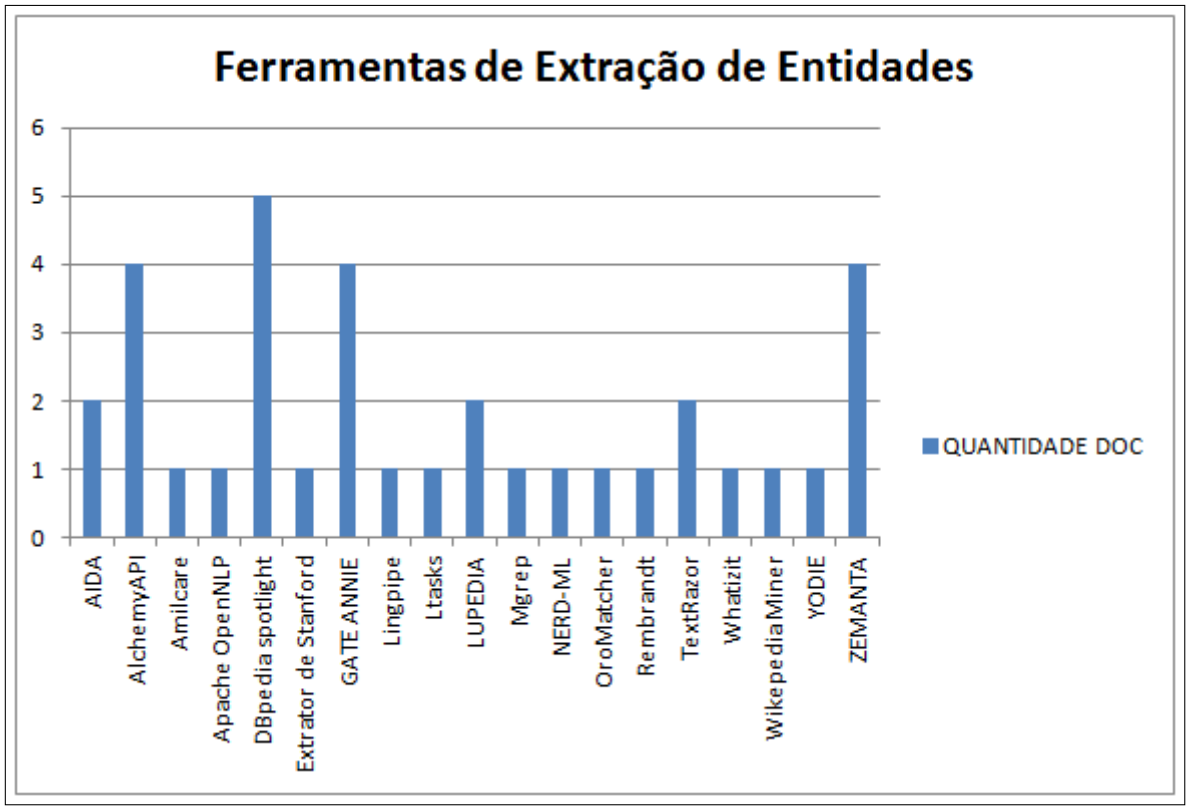
Figura 35 – Quantidade de Ferramentas de Anotação Identificada na RSL



Fonte: Próprio Autor.

Para trabalhar com extração de entidades (EE), foi identificado na revisão os extratores de entidade mencionados na figura 36 .

Figura 36 – Quantidade de Ferramentas de Extração de Entidade Identificada na RSL



Fonte: Próprio Autor.

Sendo que a ferramenta TextRazor que trabalha com as línguas inglês, francês, italiano, espanhol, russo e com o português foi empregada dentro da abordagem dessa dissertação.

R3: Dos componentes/módulos encontrados, qual(is) permite executar a ação de anotar com eficiência?

Na pesquisa realizada por (REEVE; HAN, 2005) ele comparou ferramentas para anotação semântica disponíveis até o presente ano da sua pesquisa, ressaltou suas características e apurou a eficácia de suas anotações. Ele destacou que a ferramenta MnM, que utiliza aprendizagem de máquina na identificação das entidades, como a de melhor desempenho e a de pior a Onto-O-Mat. Em suas conclusões ele informa que algoritmos de aprendizagem de máquina são mais efetivos do que os métodos baseados em padrões, porém os sistemas baseados em regras podem possuir uma performance melhor do que os sistemas baseados em aprendizagem de máquina.

Podemos concluir dessa revisão que para criar um arcabouço de anotação semântico é fundamental um componente de extração de entidade (EE) e um componente que execute o anotação no documento corrente com entidades interligadas as ontologias desejadas. Nessa revisão, observamos a escassez de extratores de entidades para a língua

portuguesa. O trabalho de (DERCZYNSKI DIANA MAYNARD, 2014) enriquece essa dissertação citando a solução para esse problema com o EE **TextRazor** e ou **Alchemy** que além dos idiomas como o inglês, espanhol, italiano, russo, alemão, também oferecem um serviço de identificação de entidade para a portuguesa.

3.5 Trabalhos Relacionados

Para trabalhos relacionados buscou-se pesquisas que estão relacionadas com estudos sobre a Web Semântica, na questão de prover semântica a documentos web utilizando um framework para alcançar anotação semântica com LOD. Esse capítulo é organizado através de um quadro com as principais características encontradas nos quatro projetos e depois uma descrição de cada um dessas pesquisas.

De acordo com a figura 37, observamos que todos os quatro trabalhos possuem como maneira de anotar os documentos o tipo automático, mas somente o trabalho do (FAFALIOS; PAPADAKOS, 2014) é que possui como origem das informações para auxiliar no processo de anotação os dados abertos conectados (LOD). O documento de entrada, aquele que é lido para ser realizado a anotação, diversificou entre os projetos, mas o documento web (Html) foi identificado na pesquisa de (FAFALIOS; PAPADAKOS, 2014) e (NETO, 2009). Para esses trabalhos temos a característica de extensibilidade, isso é: o sistema que os autores trabalharam no processo de anotação, aceita configurar e ser adaptado a novos domínios e outras ferramentas que auxiliem nesse processo.

Figura 37 – Tabela comparativa entre as características dos trabalhos relacionados

Autores	Características dos Trabalhos Relacionados						
	Ferramenta	Doc. Entrada	Origem dos Dados	Extensível	Forma de Anotação	Onde Anota	Estrutura da Anotação
(FAFALIOS; PAPADAKOS, 2014)	Theophrastus	Html, Pdf	Linked Open Data	Sim	Automática	Próprio Doc. Entrada	-
(SANTOS NETO, Gilberto Martins dos, 2009)	Semantic Web Annotation	Html	Específico	Sim	Automática	Gera outro Doc.	OWL
(FONTES; CAVALCANTI; MOURA, 2013)	Autômeta	-	Específico	-	Automática	Próprio Doc. Entrada	RDFa
(ZHANG; CHEN; FENG, 2013)	DBpedia Spotlight	Xml	Dbpedia	-	Automática	Gera outro Doc.	RDFa

Fonte: Próprio Autor.

A pesquisa desenvolvida por (FAFALIOS; PAPADAKOS, 2014), *Theophrastus: an demand and real-time automatic annotation and exploration of (web) documents using open linked data*, que suporta a anotação automática de documentos da web por meio de mineração de entidade e fornece serviços de exploração de dados utilizando o conjunto de dados abertos (LOD) em tempo real. Foi apresentado para biólogos da marinha, o sistema que atua no domínio de biodiversidade, e tem como objetivo resolver o demorado e custoso processo de identificação, desambiguar e coletar informações de espécies. É um sistema configurável e adaptável para diferentes domínios de interesse. O Theophrastus não processa RDFa ou Microdados que possam estar em uma página da web, ele identifica as entidades de interesse com base em sua configuração atual e anota as entidades detectadas no próprio documento. O trabalho apresenta ferramentas de extração de entidades baseadas no LOD (DBpedia spotlight, AlchemyAPI, Calais, AIDA e Wikimeta).

O trabalho de (NETO, 2009), *Anotação Semântica De Recursos Web Baseada em Ontologias*, mostrou o desenvolvimento de um método totalmente automático para anotar semanticamente recursos Web, em particular páginas em HTML, visando obter descrições dos termos existentes nos recursos Web. O projeto foi elaborado em três partes: de extração dos termos dos recursos, mapeamento semântico e anotação dos conceitos identificados, sendo que a ênfase é dada ao mapeamento e à anotação semântica, que tem como base a linguagem OWL. Esse projeto gerou um arcabouço de software, Semantic Web Annotation Framework, que é um sistema de componentes orientado a objetos visando generalização e reutilização. Ele trabalha com duas ontologias para os experimentos nesse trabalho: uma chamada Autos de domínio específico, relacionada a automóveis e outra de domínio genérico, chamada ontologia de topo, denominada de SUMO. Esse framework é extensível, pois, sua estrutura permite a utilização de outras ontologias, bastando que elas estejam no padrão OWL DL, e permite também a agregação de outros extratores para formatos diferentes de HTML e outras ferramentas de mapeamento semântico que empreguem técnicas diferentes do padrão adotado. Nos trabalhos relacionados na pesquisa de (NETO, 2009), ele destacou as ferramentas Amilcare, SemTag, Seeker, Ont-O-Mat, MnM e WEESA. Essas três ferramentas apresentam uma desvantagem por serem métodos de anotação semi-automáticos, pois precisam de fases de treinamento e supervisão humana ao contrário do projeto criado por ele. Dentre as questões que esse autor relata sobre os métodos para a anotação semântica automática SemTag, Seeker e KIM, pode se dizer que o problema delas é a utilização de apenas uma ontologia, inviabilizando a aplicação que o conteúdo dos recursos Web seja de um domínio diferente do domínio da ontologia.

O trabalho de (FONTES; CAVALCANTI; MOURA, 2013), *An Ontology-Based Reasoning Approach for Document Annotation*, apresenta uma proposta para enriquecer automaticamente documentos com anotações semânticas, a anotação é realizada de acordo com uma ontologia de domínio. Além de transformar documentos semânticos, inclui a noção de meta-anotação (anotação sobre anotação). Nessa pesquisa são mencionadas outras

ferramentas como: Zemanta, Annotea, GATE, SMORE, OpenCalais e GoNTogle. Diferente dessas ferramentas que não exploram intensamente o potencial de inferência de ontologias, ele apresenta como resultado do seu trabalho a ferramenta Autômeta, que tem a funcionalidade de inferir sobre uma ontologia de domínio específico (não trabalha com técnicas de processamento de linguagem natural) e gerar documentos anotados de forma intrusiva no formato RDFa.

O trabalho de (ZHANG; CHEN; FENG, 2013), com o texto *Semantic Annotation for Web Services Based on DBpedia*, propõe anotação semântica com base no DBpedia. O enriquecimento é realizado por meio do conjunto de dados abertos interligados (LOD), ontologias do DBpedia. Ferramentas são mencionadas (METEOR-S e ASSAM). Para o seu processo de anotação semântica é utilizado DBpedia Spotlight e o Domj4.

Os trabalhos mencionados são os pilares para o desenvolvimento desta dissertação que possui grande semelhança com os primeiros trabalho mencionados. A nossa pesquisa desenvolveu-se de forma semelhante aos trabalhos de (FAFALIOS; PAPADAKOS, 2014), (NETO, 2009). Como na pesquisa desses autores, pode-se reutilizar nesta dissertação a ideia do Extrator de Informação, Mapeador Semântico e a base de dados aberta (Linked Open Data) comentada pelo autor (ZHANG; CHEN; FENG, 2013). E se necessário, reutilizar as ferramentas mencionadas para alcançar o objetivo de anotar os documentos do currículo Lattes.

4 Metodologia

Conforme (GIL, 2002), a pesquisa experimental consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e observação dos efeitos que a variável produz no objeto. Ela caracteriza-se por manusear sem desvios as variantes associadas ao núcleo de estudo. De acordo com (CERVO; BERVIAN; SILVA, 2006), a experimentação é um conjunto de processos usados para conferir as hipóteses, sendo essa uma relação de causa e efeito ou de antecedência e consequência entre os dois acontecimentos.

A natureza da pesquisa é aplicada, porque há um interesse em adquirir conhecimentos orientados para a aplicação prática. Essa natureza de pesquisa é realizada para determinar as possíveis utilidades para as descobertas da pesquisa ou definir novos métodos ou maneiras de alcançar a solução de problemas específicos (CASARIN; CASARIN, 2011). Apesar da natureza da pesquisa ser do tipo aplicada, ligada à prática, essa não pode deixar de incluir consideração e pensamentos teóricos(MASCARENHAS,).

Com relação ao problema da pesquisa, ele é qualitativo, pois prevalece a parte descritiva, onde os objetivos envolvem a descrição de um fenômeno, caracterizando sua ocorrência e relacionando-o com outros fatores (CASARIN; CASARIN, 2011). Segundo (MASCARENHAS,) a pesquisa qualitativa é feita quando deseja-se descrever com detalhes um objeto de estudo. Neste âmbito, a pesquisa será explicativa, pois procura identificar fatores que determinam ou contribuem para a ocorrência dos fenômenos (GIL, 2002).

Por intermediação das definições de tipologias de pesquisas apresentadas, pode-se afirmar que a pesquisa do tipo prova de conceito e experimento é um caminho adequado para o desenvolvimento dessa pesquisa. Em pesquisas explicativas, o método utilizado é o experimental, sendo predominante na área de ciências exatas (CASARIN; CASARIN, 2011).

Figura 38 – Relação entre: Atividades Metodológicas X Objetivos Específicos

Atividades Metológicas	**OBJETIVOS ESPECÍFICOS**		
	Conceituar e identificar as tecnologias relacionadas com Anotação Semântica.	Selecionar os currículos dos docentes da plataforma Lattes.	Anotar o currículo Lattes.
Revisão Sistemática da Literatura			
Extrair os currículos da Plataforma Lattes.			
Identificação dos DataSets disponíveis e a forma de interligá-los.			
Identificação e seleção dos softwares necessários para anotar do Lattes.			
Arcabouço de anotação do Lattes.			

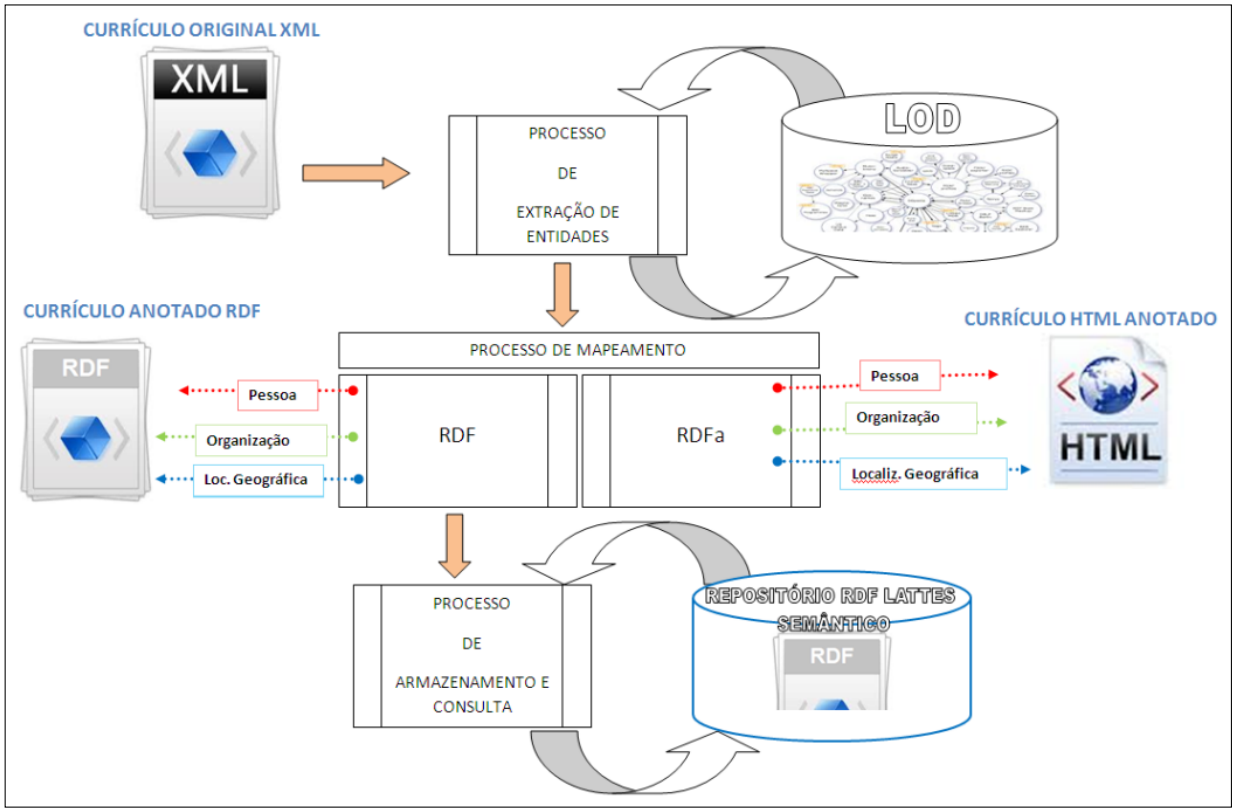
Os procedimentos metodológicos para atingir os objetivos específicos estão abordados na figura 38. O trabalho foi desenvolvido em duas partes: a primeira onde constatá-se do levantamento bibliográfico sobre Anotação Semântica, abrangendo conceitos, componentes, ferramentas, tecnologias e suas aplicações, com o objetivo de auxiliar na implementação do arcabouço. Na segunda parte, foi criado de acordo com o arcabouço a aplicabilidade dos conceitos, ferramentas e tecnologias para a execução da Anotação Semântica. O arcabouço conceitual é importante porque representa um modelo da arquitetura necessária para implementar anotações automáticas utilizando Linked Open Data através de qualquer técnica.

5 Arcabouço Conceitual

O objetivo deste trabalho será anotar automaticamente os documentos web do currículo Lattes utilizando as ontologias disponíveis e as tecnologias de anotação semântica.

A figura 39 representa a proposta implementada da arquitetura conceitual de anotação semântica e tem como objetivo identificar os componentes, suas relações e etapas que farão parte desta dissertação.

Figura 39 – Modelo conceitual do projeto



Fonte: Próprio Autor.

Podemos explicar o processo nas seguintes etapas:

1. Os dados que alimentam o sistema são documentos/currículos Lattes no formato XML.

2. O Processo de Extração tem a função de extrair os termos do currículo, utilizando o componente externo denominado de TextRazor. Ele é uma ferramenta de extração que presta o serviço de extração de entidade via web, um web service, possui uma

abordagem aprendizagem de máquina (NLP), de domíno geral, interligada com as bases abertas do DBPedia e Freebase, porém a licença de uso não é gratuita, nessas condições a quantidade de dados a serem utilizados por mês é restrito. O resultado do processo dessa etapa é enviado para o Processo de Mapeamento.

3. O Mapeamento é realizado de maneira que possibilita identificar entidades de um domínio de interesse no currículo. O componente de mapeamento semântico foi desenvolvido utilizando os dados gerados pelo TextRazor, afim de gerar um documento em uma estrutura RDFa e RDF Turtle. Essas estruturas permitem que o documento seja entendido tanto pelas máquinas quanto pelas pessoas e ainda, realização de consultas SPARQL. A geração do documento RDF Turtle, com as triplas de anotações identificadas no Currículo Lattes XML, seguiu o padrão do vocabulário Anotação Aberta (OA)[1].

4. No componente, Armazenamento e Consulta, é realizado o armazenamento do documento RDF de triplas em um banco de dados (tripleStore) que fornece um modo padrão para compartilhamento de dados, intercâmbio, permite consultar e manipular os dados usando a linguagem de consulta SPARQL (TAO et al., 2013).

No fim do processo temos o currículo Lattes anotado com os dados semânticos, estará legível para as pessoas e máquinas, abrindo possibilidades para os motores de busca inteligentes e a realização de consultas sofisticadas. Esse modelo foi elaborado a partir das informações e ferramentas levantadas na RSL.

[1] Open Annotation Data Model- é uma estrutura para a criação de associações entre recursos relacionados que está de acordo com a arquitetura da World Wide Web

6 Implementação

A implementação do sistema de anotação semântica, denominado de Lattes Web Semântico[1], iniciou-se a partir da definição do programa de extração de entidade na parte da Revisão Sistemática.

A figura 40 representa a visão geral dos componentes do sistema. Esse sistema executa a funcionalidade de mapeamento semântico com a geração dos arquivos HTML(RDFa) e RDF com as entidades localizadas e anotadas baseadas no LOD.

Figura 40 – Visão Geral dos Componentes do Sistema Lattes Web Semântico

Fonte: Próprio Autor.

O Sistema Lattes Web Semântico (LattesWS) foi desenvolvido na linguagem PHP, auxiliado por um Web Service para a extração de entidade e um gerenciador de triplas RDF. A imagem 41 representa uma visão geral da interação entre os componentes que integram esse Sistema.

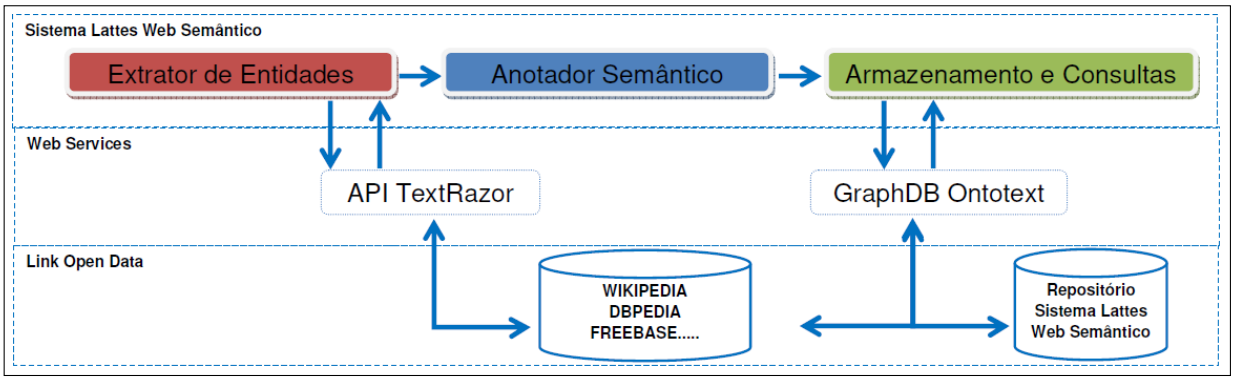
O sistema é composto pelas seguintes unidades: Extrator de Entidades, Anotador Semântico (Anotador Semântico RDFa e Anotador Semântico RDF), Armazenamento das Triplas e Consulta Semântica.

6.1 Módulo de Extração de Entidades

Esse componente tem como objetivo identificar as entidades, seus tipos e links dos bancos de dados abertos em um texto. Para essa tarefa foi definida a ferramenta TextRazor, identificada nas pesquisas realizadas na RSL deste trabalho. Apesar dessa ser um serviço comercial, ela trabalha com reconhecimento de entidades utilizando aprendizagem de máquina, reconhece os textos em português e retorna links das bases abertas DBPedia e

[1] www.wssistemas.com.br/latteswss/

Figura 41 – Visão Geral da Iteração entre os componentes do Sistema Lattes Web Semaântico



Fonte: Próprio Autor.

Freebase. A API TextRazor possibilita a extração de entidades, desambiguação, conexões de bases abertas e classificação em 10 tipos de língua. Na sua versão gratuita, empreendida neste trabalho, é possível efetuar 500 análise de texto por dia com um tamanho de 10KB cada texto (TEXTRAZOR, 2015).

Esse módulo foi construído com a linguagem de desenvolvimento em PHP com o auxílio do TextRazor para trabalhar com o processamento de extração de entidade dos arquivos Lattes no formato XML. Esse processamento compreende: a identificação dos arquivos que não foram processados em um diretório; envio e recebimento de dados para o web service de extração; tratamento e preparação dos dados recebidos do web service para posterior mapeamento.

Na identificação dos arquivos Lattes, o módulo verifica se existem arquivos disponíveis no diretório para processamento. Caso exista, envia as partes desejadas do arquivo XML (dados do responsável, resumo, endereço, formação acadêmica e áreas de conhecimento) para o web service de extração de entidade, TextRazor. Esse extrator recebe os textos e retorna: um objeto contendo o termo ou palavra identificada, o tipo de entidade (domínio que o termo pertence: Organização, Pessoa, País, Estado, Cidade, Universidade...) e o link que representa esse termo no Wikipedia, Dbpedia e ou Freebase. No final desse módulo é realizado o tratamento dessa resposta: verificando e organizando as entidades localizadas e não localizadas, a identificação do tipo que a entidade pertence e seus links para as bases abertas (Dbpedia e Freebase).

6.2 Módulo de Anotação Semântica das Entidades

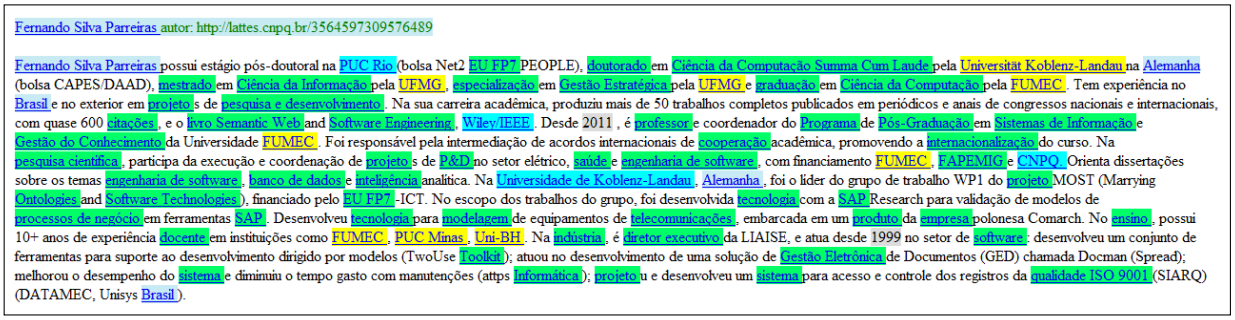
6.2.1 Anotação em RDFa

Esse componente tem como objetivo receber os dados gerados pelo TextRazor, manipulá-los e gerar um arquivo HTML com uma estrutura RDFa embutida no mesmo arquivo.

Com o texto que representa uma parte do arquivo Lattes XML (dados do responsável, resumo, endereço, formação acadêmica e áreas de conhecimento) e a resposta do TextRazor para cada uma dessas partes, seleciona-se os termos anotados e o tipo de entidade de cada termo para criar uma formatação RDFa. A formatação RDFa é gerada para cada tipo de entidade (Pessoa, Organização, Universidade, País, Ano..) especificando o seu domínio e propriedades definidas previamente na etapa de desenvolvimento.

A figura 42 representa um exemplo do arquivo HTML e as anotações no formato RDFa para o campo **resumo** do arquivo XML do Lattes:

Figura 42 – Exemplo de Anotação Semântica do Resumo do Currículo Lattes no Formato RDFa



Fonte: Próprio Autor.

Na imagem 42, temos que cada tipo de entidade (Pessoa, Empresa, Organização, País, Universidade, Ano e Coisa[Entidade Geral]) é identificada por uma cor. Internamente, de acordo com os exemplos das imagens 43,44,45, temos o código da página HTML com as anotações RDFa embutida para cada um desses tipos:

Figura 43 – Exemplo 1: de Anotação Semântica, entidade Pessoa, do Resumo do Currículo Lattes

```
<div class="rdfa_pessoa" typeof="foaf:Person schema:Person">
      <a property="url" href="">
            <span property="foaf:name">Fernando Silva Parreiras</span>
      </a></div>
```

Fonte: Próprio Autor.

Figura 44 – Exemplo 2: de Anotação Semântica, entidade País, do Resumo do Currículo Lattes

```
<div class="rdfa_Pais" typeof="schema:Country">
    <a property="url" href="http://pt.wikipedia.org/wiki/Alemanha">
        <span property="schema:name"> Alemanha</span>
   </a>
   <link property="schema:sameAs" href="http://pt.wikipedia.org/wiki/Alemanha"/>
</div>
```

Fonte: Próprio Autor.

Figura 45 – Exemplo 3: de Anotação Semântica, entidade Coisas, do Resumo do Currículo Lattes

```
<div class="rdfa_Coisa" typeof="schema:Thing">
    <a property="url" href="http://pt.wikipedia.org/wiki/Especialização">
        <span property="schema:name">especialização</span>
    </a>
    <link property="schema:sameAs" href="http://pt.wikipedia.org/wiki/Especialização"/>
</div>
```

Fonte: Próprio Autor.

A anotação semântica é realizada com o código RDFa embutido em códigos HTML. No caso de entidades do tipo País, a anotação está sendo realizada como no exemplo da imagem 44. Entidades do tipo Pessoa, figura 43, são mapeadas com os vocabulários FOAF e SCHEMA:PERSON utilizando as propriedades URL e NAME.

Se o extrator de entidade não identificar o tipo de uma entidade, como na figura 45, mas se houver a identificação do termo ou palavra e o conceito em uma base aberta, LOD, a entidade é mapeada como "Coisa"(Thing) a partir do vocabulário SCHEMA:Thing utilizando a propriedade NAME e SAMEAS.

Essa forma de codificar, RDFa, foi definida por trabalhar concomitante com os códigos HTML, ficando transparente para o usuário e proporcionando páginas ou documentos com a semântica embutida.

6.2.2 Anotação em RDF

O componente de anotação semântica em RDF tem como objetivo receber os dados gerados pelo TextRazor, manipulá-los e gerar um arquivo Turtle(ttl). O arquivo gerado é inserido em um sistema de banco de dados de triplas, o que permite efetuarmos as consultas semânticas.

A lógica desse módulo é semelhante ao módulo anterior: com as partes do texto arquivo Lattes XML (dados do responsável, resumo, endereço, formação acadêmica e áreas de conhecimento) e a resposta do TextRazor para cada uma dessas partes, seleciona-se o termo ou palavra e seu tipo de entidade que foi localizada para criar um arquivo Turtle de triplas RDF (exemplo na figura 46). A formatação das triplas RDF é gerada para cada tipo de entidade (Pessoa, Organização, Universidade, País, Ano..) especificando o seu domínio e propriedades previamente definidas.

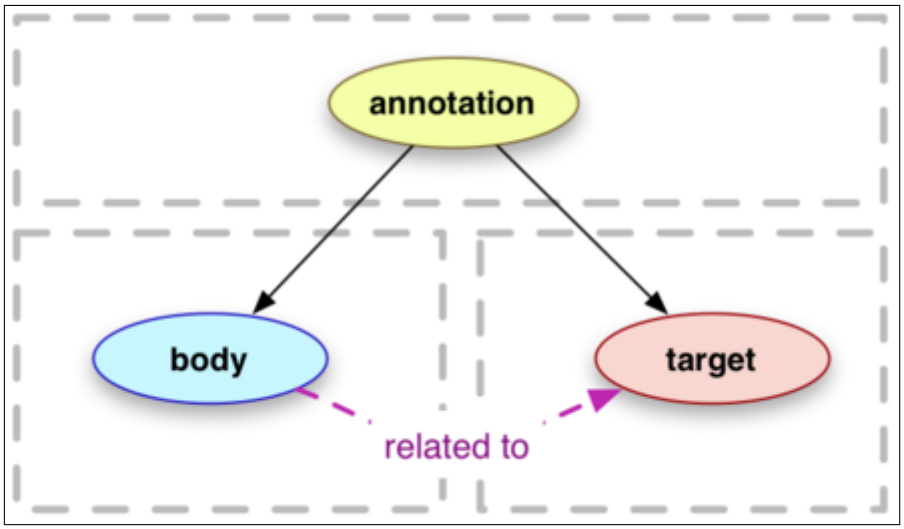
Figura 46 – Exemplo Fragmento de um RDF

```
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc:     <http://purl.org/dc/terms/> .
@prefix schema: <http://schema.org/>.
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .
@prefix cv:<http://rdfs.org/resume-rdf/cv.rdfs#>.
@prefix oa:     <http://www.w3.org/ns/oa#>.
@prefix cnt:    <http://www.w3.org/2011/content#> .
@prefix prov:   <http://www.w3.org/ns/prov#> .
@prefix lattes: <http://lattes.cnpq.br/>.
@prefix wss:    <http://www.wssistemas.com.br/latteswss/>.
@prefix wss_rdfa:   <http://www.wssistemas.com.br/latteswss/recursordfa/>.

<wss_rdfa:3564597309576489> rdf:type        cv:CV;
            cv:aboutPerson  "Fernando Silva Parreiras";
            cv:cvTitle      "Currículo Lattes Anotado em RDFa do Sistema de Currículos Lattes";
            dc:creator      "Fernando Silva Parreiras";
            schema:sameAs   <lattes:3564597309576489>;
            dc:format   "text/html".
```

Fonte:Próprio Autor.

Os dados provenientes do TextRazor são manipulados por esse módulo com o intuito de criar as triplas utilizando vocabulários e um modelo de anotação de dados abertos. Isso é, as anotações descritas nesse arquivo (ttl) segue a representação do Modelo de Anotação de Dados Abertos (Open Annotation Data Model - OA) que é um estrutura, framework, para a criação de associações entre recursos relacionados, anotações usando uma metodologia que está de acordo com a arquitetura da World Wide Web. Uma anotação é considerada um conjunto de recursos conectados que inclui uma parte principal e um destino secundário. No qual o item principal transmite algo sobre o item secundário. Em seu nível elementar, o modelo OA diferencia entre três componentes: anotação, itens e destino, onde a anotação expressa de que um item está relacionado com o destino da anotação (Figura 47), (GROUP, 2013).

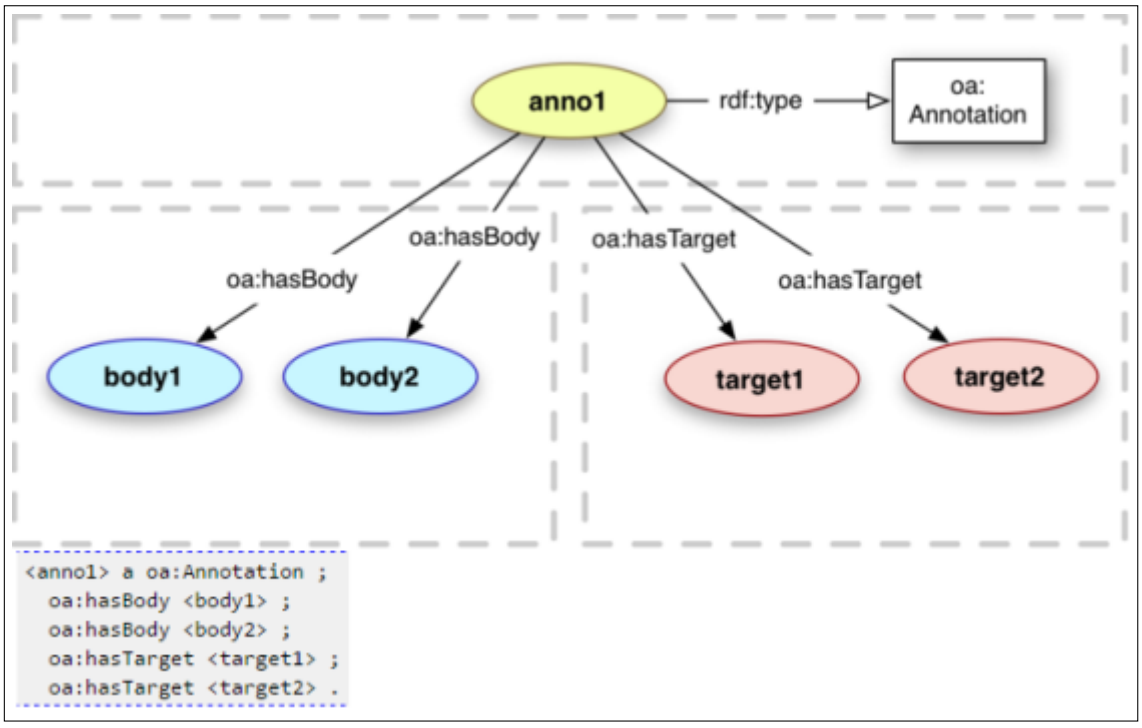
Figura 47 – Open Annotation Data Model: Annotation, Body and Target



Fonte:(GROUP, 2013)

Esse modelo disponibiliza uma maneira flexível para a criação das anotações, independente da quantidade de itens, possibilitando criar diversos itens e destinos. A figura 48 representa o modelo dessa situação, onde cada "Body"é igualmente relacionado com seu "Target".

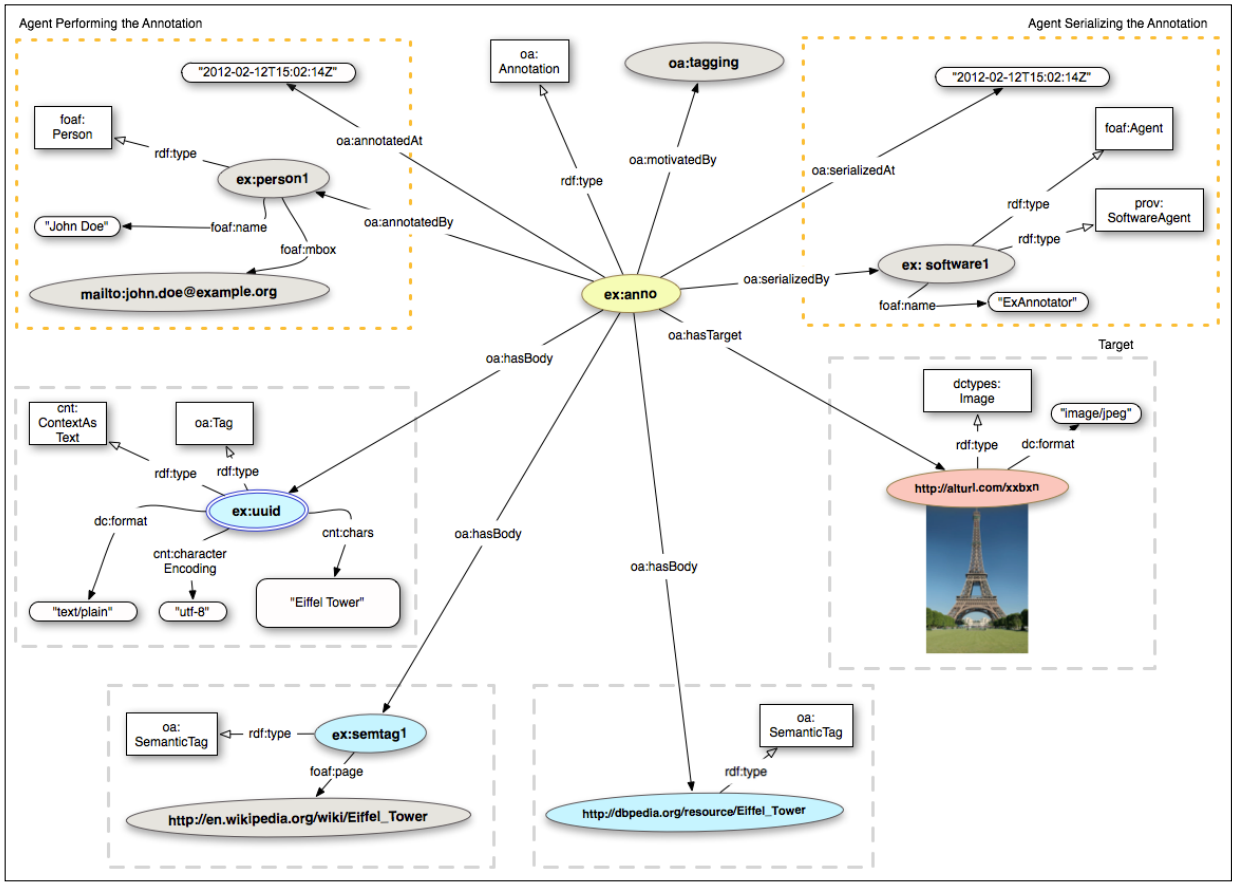
Figura 48 – Open Annotation Data Model: Diversos corpos ou objetivos



Fonte:(GROUP, 2013)

A flexibilidade desse modelo permite mapearmos vários itens para a mesma entidade. A figura 49 modela a reunião das três maneiras diferentes de marcas de codificação em uma única anotação. A existência dessa forma de marcação, três marcas diferentes para a mesma entidade, é utilizada com o objetivo de inferir mapeamentos.

Figura 49 – Open Annotation Data Model: Multiple Tags



Fonte:(GROUP, 2013)

A representação Turtle (ttl) para a figura 49 pode ser observada na figura 50.

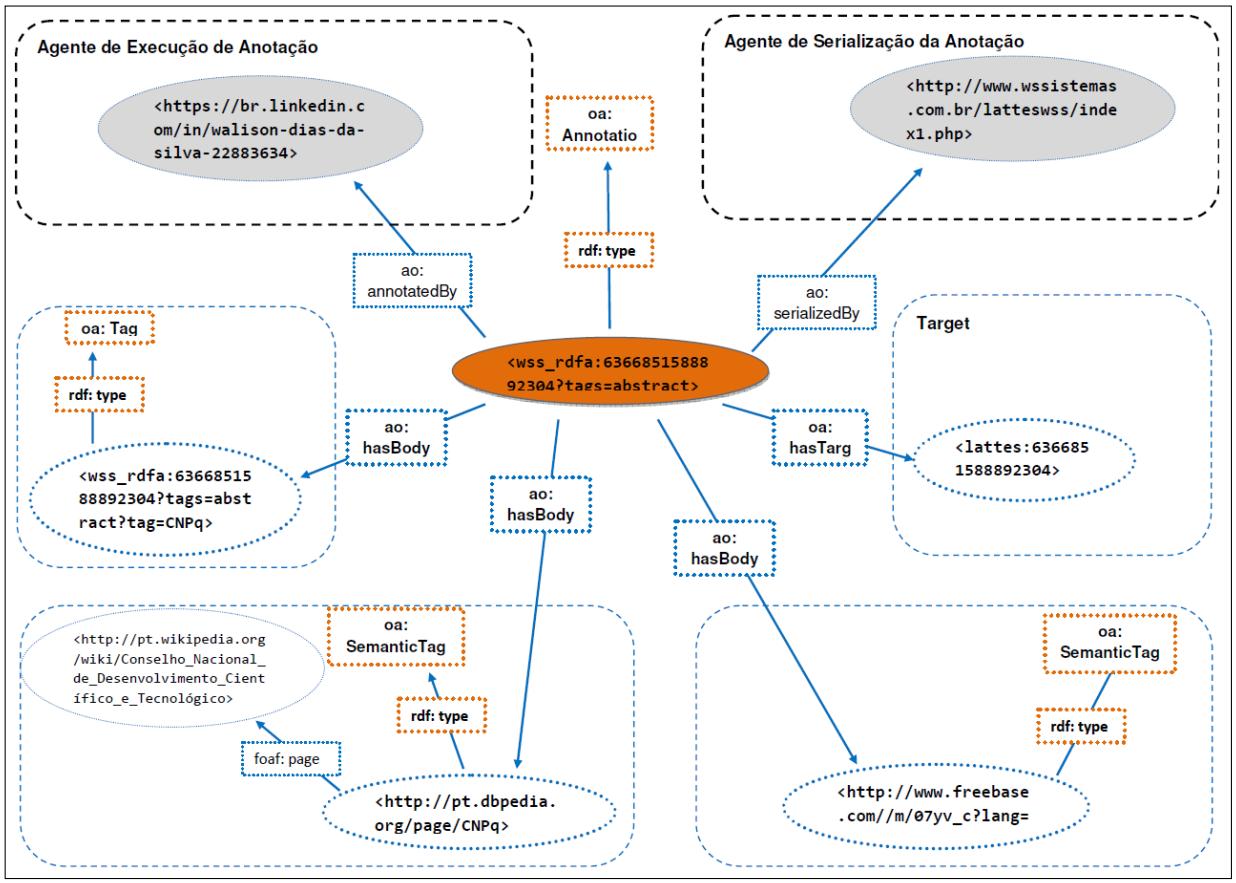
Figura 50 – Open Annotation Data Model: RDF Turtle para Multiplas Tags

```
ex:anno a oa:Annotation ;
  oa:hasTarget <http://alturl.com/xxbxn> ;
  oa:hasBody ex:uuid ;
  oa:hasBody ex:semtag1 ;
  oa:hasBody <http://dbpedia.org/resource/Eiffel_Tower>;
  oa:motivatedBy oa:tagging ;
  oa:annotatedBy ex:person1 ;
  oa:annotatedAt "2012-02-12T15:02:14Z" ;
  oa:serializedBy ex:software1 ;
  oa:serializedAt "2012-02-12T15:02:14Z" .

<http://alturl.com/xxbxn> a dctypes:Image
  dc:format "image/jpeg" .

ex:uuid a cnt:ContentAsText ;
  cnt:chars "Eiffel Tower" ;
  dc:format "text/plain" ;
  cnt:characterEncoding "utf-8" .

ex:semtag1 a oa:SemanticTag ;
  foaf:page <http://en.wikipedia.org/wiki/Eiffel_Tower> .

<http://dbpedia.org/resource/Eiffel_Tower> a oa:SemanticTag.

ex:person1 a foaf:Person ;
  foaf:mbox <mailto:john.doe@example.org> ;
  foaf:name "John Doe" .

ex:software1 a foaf:Agent, prov:SoftwareAgent ;
  foaf:name "ExAnnotator" .
```

Fonte:(GROUP, 2013)

Então, seguindo os padrões e com os exemplos do modelo OA conclui-se e cria-se a estrutura do arquivo Turtle RDF do Sistema Lattes Semântico com a representação dos termos/palavras encontradas nos arquivos Lattes XML de acordo com o modelo presente na figura 51.

Figura 51 – Modelo de Anotação Semântica dos Arquivos Turtle RDF do Sistema Lattes Web Semântico



Fonte:Próprio Autor.

Existem ferramentas que possibilitam gerar automáticamente a partir de um texto RDFa um texto RDF, porém neste trabalho desenvolve-se o módulo que cria o arquivo RDF afim de evitar dependências de sistemas externos e por criar um RDF de acordo com os padrões de OA.

6.3 Módulo de Armazenamento e Consultas

Foi definido para armazenar e gerenciar os arquivos RDF Turtle um banco de dados que armazena triplas RDF e possibilita efetuar alterações e consulta semânticas sobre os dados armazenados.

O GraphDB é um repositório de dados semânticos usado para o armazenamento, consulta e gerenciamento de dados estruturados, utilizando ontologias para raciocinar automaticamente sobre os dados. De acordo com o (ONTOTEXT, 2015), a versão gratuita permite armazenar no máximo 100 milhões de triplas, executa consultas SPARQL 1.1,

e suporta operações de raciocínio para inferência. O GraphDB S4 é a solução ideal para cenários com o tamanho do banco de dados pequeno ou médio e alta carga de consulta, onde investir em licenças de software, provisionamento e manutenção de um servidor é um custo não ideal. Com essa solução efetua-se de forma manual a inserção das triplas através do upload do arquivo RDF Turtle ou executa-se inserções automáticas com os dados do documento RDF.

As consultas do Sistema Lattes Web Semântico são previamente definidas. Para cada consulta, desenvolve-se na linguagem PHP o acesso, execução e tratamento do retorno do resultado da consulta para os usuários. As consultas são realizadas na linguagem de consulta estruturada SPARQL, padrão da Web Semântica.

7 Validação do Sistema

Esse capítulo tem como intenção demonstrar o que a aplicação dos conceito pesquisados na RSL e as definições do modelo de arcabouço permitiu construir. A efetividade do sistema acontece a partir da existência da anotação dos arquivos gerados no formato RDFa, nas consultas geradas sobre os dados RDF Turtle armazenados e na interligação dos dados com LOD.

Como mencionando anteriormente, as pessoas observam as anotações realizadas no arquivo Html/RDFa. Na figura 52 é representado um trecho de um documento Lattes anotado pelo Sistema Lattes Web Semântico. Nas linhas 16 a 20 é acrescentado pelo componente de anotação semântica o criador/proprietário das informações do documento. Nas linhas 21 a 25 é identificado pelo extrator de entidade (EE) que o nome (Fernando Silva Parreiras) é uma Pessoa, então foi aplicada uma essa estrutura de marcação de tipos e propriedades para o termo identificado como uma entidade do tipo Pessoa (Person). O termo (PUC Rio) foi outra palavra identificada, linhas 26 a 28, onde o EE informou que essa é uma Organização, nesse contexto foi elaborada uma estrutura de marcação para esse tipo de entidade.

Figura 52 – Anotação Semântica do Currículo Lattes no Sistema Lattes Web Semântico

```
16  <div class="AutorLattes" resource="http://lattes.cnpq.br/" typeof="">
17      <a property="url" href="http://lattes.cnpq.br/">
18      <span property="dc:creator">Fernando Silva Parreiras</span>
19      </a> autor: http://lattes.cnpq.br/3564597309576489
20  </div>
21  <div class="rdfa_pessoa" typeof="foaf:Person schema:Person">
22    <a property="url" href="">
23          <span property="foaf:name">Fernando Silva Parreiras</span>
24    </a>
25  </div> possui estágio pós-doutoral na
26  <div class="rdfa_Company" resource="" typeof="schema:Company">
27    <span property="schema::name"> PUC Rio</span>
28  </div> ((bolsa Net2
29  <div class="rdfa_Coisa" typeof="schema:Thing">
30      <a property="url" href="http://en.wikipedia.org/wiki/Framework_Programmes_for_Research_and_
31          <span property="schema:name">EU FP7</span>
32      </a>
33      <link property="schema:sameAs" href="http://en.wikipedia.org/wiki/Framework_Programmes_for_
34  </div>  PEOPLE),
35  <div class="rdfa_Coisa" typeof="schema:Thing">
36      <a property="url" href="http://pt.wikipedia.org/wiki/Doutoramento">
37          <span property="schema:name">doutorado</span>
```

Fonte:Próprio Autor.

Esses arquivos são armazenados no diretório e podem ser observados no próprio site da aplicação. As figuras 53 e 54 representam a visão da anotação do Resumo de um documento Lattes no formato RDFa e RDF Turtle consecutivamente disponíveis no Sistema LattesWS.

Figura 53 – Resumo do Currículo Lattes Anotado no Sistema LattesWS em RDFa

Fonte:Próprio Autor.

Figura 54 – Resumo do Currículo Lattes Anotado no Sistema LattesWS em RDF Turtle

```
<wss_rdfa:3564597309576489?tags=abstract>   a   oa:Annotation, oa:d4e434 ;
        oa:hasTarget    <wss:/recursoxml/3564597309576489?tags=abstract>,<lattes:3564597309576489>;
        oa:annotatedBy  <http://www.wssistemas.com.br/latteswss/index1.php>,<
        https://br.linkedin.com/in/walison-dias-da-silva-22883634>;
        oa:annotatedAt  "2016-03-21T17:50:26Z" ;
        oa:serializedBy <http://www.wssistemas.com.br/latteswss/index1.php>,<https://www.textrazor.com/>;
        oa:serializedAt "2016-03-21T17:50:26Z" ;

#####HasBody do abstract:
        oa:hasBody <wss_rdfa:3564597309576489?tags=abstract?tag=CNPQ>;
            oa:hasBody <http://pt.dbpedia.org/resource/CNPQ>;
            oa:hasBody <
            http://dbpedia.org/resource/National_Council_for_Scientific_and_Technological_Development>;
            oa:hasBody <http://www.freebase.com//m/07yv_c?lang=pt>;
        oa:hasBody <wss_rdfa:3564597309576489?tags=abstract?tag=FAPEMIG>;
            oa:hasBody <http://pt.dbpedia.org/resource/FAPEMIG>;
        oa:hasBody <wss_rdfa:3564597309576489?tags=abstract?tag=Universidade_FUMEC>;
            oa:hasBody <http://pt.dbpedia.org/resource/Universidade_FUMEC>;
            oa:hasBody <http://dbpedia.org/resource/FUMEC_University>;
            oa:hasBody <http://www.freebase.com//m/0124rd0p?lang=pt>;
```

Fonte:Próprio Autor.

Outra parte relevante para essa seção de validação são as consultas. As consultas podem ser realizadas utilizando o EndPoint disponibilizado pelo sistema do banco de dados de triplas, GraphDB, ou pelo próprio sistema desenvolvido. Como o intuito nesse momento é mostrar a efetividade do que as anotações semânticas do Lattes podem oferecer, as imagens a seguir foram geradas no EndPoint da aplicação do GraphDB, porém as mesmas estão disponíveis no Sistema Lattes Web Semântico.

A figura 55 representa a consulta e o resultado da primeira pergunta: Quais são os currículos anotados e cadastrados no sistema?

Figura 55 – Consulta 01 no Sistema Lattes Web Semântico

```
195  PREFIX cv: <http://rdfs.org/resume-rdf/cv.rdfs>
196  SELECT ?nome    ?currLatt
197  WHERE {
198      ?currLatt a cv:CV ;
199          cv:aboutPerson  ?nome.
200  }
201  order by ?nome
```

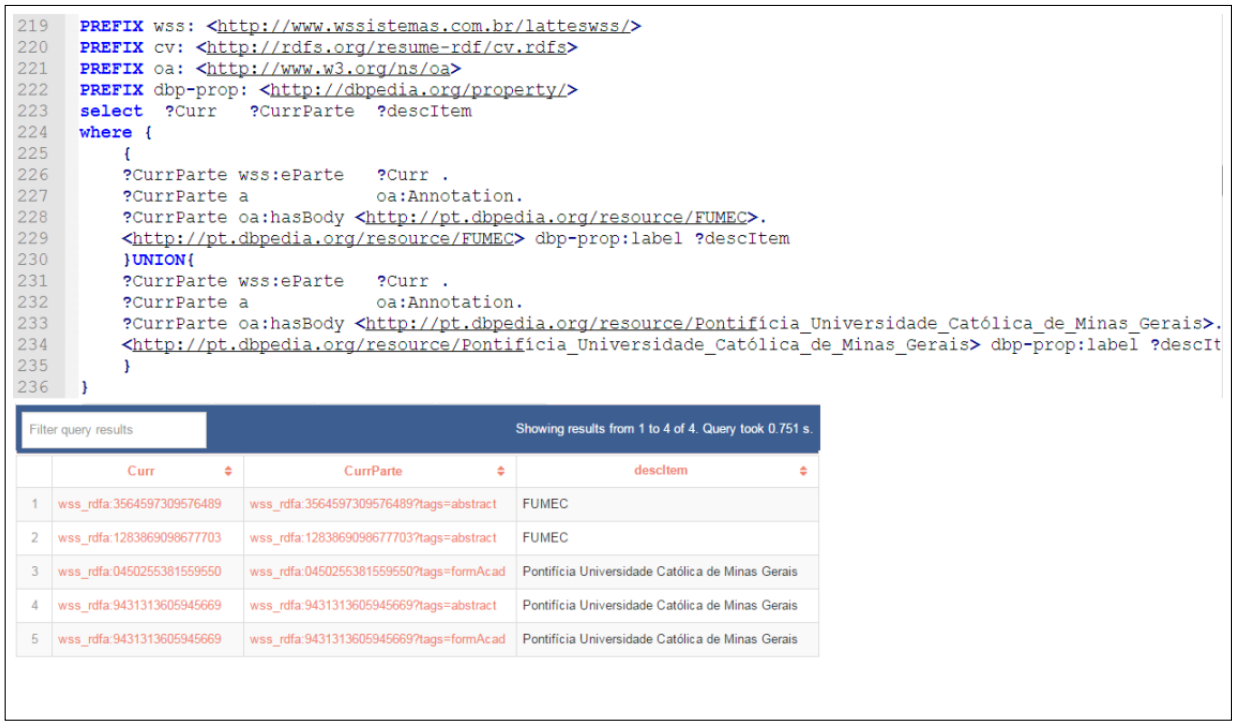
Filter query results — Showing results from 1 to 13 of 13. Query took 0.209 s.

	nome	currLatt
1	Anselmo Maciel Nunes	wss_rdfa:9408286832772730
2	Carla Geovana do Nascimento Macario	wss_rdfa:6366851588892304
3	Cristiana Fernandes De Muylder	wss_rdfa:0450255381559550
4	Eduardo Neves Motta	wss_rdfa:6726713901235343
5	Elizabeth Almeida Rolim	wss_rdfa:8042937271101537
6	Fabricio Ziviani	wss_rdfa:1283869098677703
7	Fernando Silva Parreiras	wss_rdfa:3564597309576489
8	Jersone Tasso Moreira Silva	wss_rdfa:9431313605945669
9	Jorge Tadeu de Ramos Neves	wss_rdfa:7209599572627626
10	Luiz Claudio Gomes Maia	wss_rdfa:6502942873335887
11	Marta Macedo Kerr Pinheiro	wss_rdfa:9006683778296973
12	Rodrigo Moreno Marques	wss_rdfa:4390865555343440
13	Walison Dias da Silva	wss_rdfa:6015015722771347

Fonte:Próprio Autor.

A figura 56 representa a consulta e o resultado da segunda pergunta: Em qual parte dos currículos Lattes existem o termo FUMEC e PUC Rio? Sendo que esses termos não são procurados como uma palavra chave "FUMEC"ou "PUC Rio", mas pela sua referência semântica, nesse exemplo a do DBpedia.

Figura 56 – Consulta 02 no Sistema Lattes Web Semântico



```
PREFIX wss: <http://www.wssistemas.com.br/latteswss/>
PREFIX cv: <http://rdfs.org/resume-rdf/cv.rdfs>
PREFIX oa: <http://www.w3.org/ns/oa>
PREFIX dbp-prop: <http://dbpedia.org/property/>
select  ?Curr   ?CurrParte  ?descItem
where {
    {
    ?CurrParte wss:eParte   ?Curr .
    ?CurrParte a            oa:Annotation.
    ?CurrParte oa:hasBody <http://pt.dbpedia.org/resource/FUMEC>.
    <http://pt.dbpedia.org/resource/FUMEC> dbp-prop:label ?descItem
    }UNION{
    ?CurrParte wss:eParte   ?Curr .
    ?CurrParte a            oa:Annotation.
    ?CurrParte oa:hasBody <http://pt.dbpedia.org/resource/Pontifícia_Universidade_Católica_de_Minas_Gerais>.
    <http://pt.dbpedia.org/resource/Pontifícia_Universidade_Católica_de_Minas_Gerais> dbp-prop:label ?descIt
    }
}
```

Filter query results — Showing results from 1 to 4 of 4. Query took 0.751 s.

	Curr	CurrParte	descItem
1	wss_rdfa:3564597309576489	wss_rdfa:3564597309576489?tags=abstract	FUMEC
2	wss_rdfa:1283869098677703	wss_rdfa:1283869098677703?tags=abstract	FUMEC
3	wss_rdfa:0450255381559550	wss_rdfa:0450255381559550?tags=formAcad	Pontifícia Universidade Católica de Minas Gerais
4	wss_rdfa:9431313605945669	wss_rdfa:9431313605945669?tags=abstract	Pontifícia Universidade Católica de Minas Gerais
5	wss_rdfa:9431313605945669	wss_rdfa:9431313605945669?tags=formAcad	Pontifícia Universidade Católica de Minas Gerais

Fonte:Próprio Autor.

A imagem 57 responde a consulta: Quais são os documentos que possuem na sua Área de Formação Acadêmica o termo Engenharia Elétrica e Banco de Dados ou no seu Resumo o termo Tomada de Decisão? A ideia dessa pesquisa é apresentar a possibilidade e forma de efetuar uma pesquisa de termos semânticos dentro de áreas/partes do currículo lattes semântico.

Figura 57 – Consulta 03 no Sistema Lattes Web Semântico

```
PREFIX oa: <http://www.w3.org/ns/oa>
PREFIX wss: <http://www.wssistemas.com.br/latteswss/>
PREFIX dbp-prop: <http://dbpedia.org/property/>
select *
where  {
    {
        ?currPart   wss:pertence    "formAcad".
        ?currPart   oa:hasBody      <http://pt.dbpedia.org/resource/Engenharia_Elétrica>.
        <http://pt.dbpedia.org/resource/Engenharia_Elétrica>    dbp-prop:label ?desc1
    }UNION{
        ?currPart   wss:pertence    "formAcad".
        ?currPart   oa:hasBody      <http://pt.dbpedia.org/resource/Banco_de_Dados>.
        <http://pt.dbpedia.org/resource/Banco_de_Dados> dbp-prop:label ?desc1
    }UNION{
        ?currPart   wss:pertence    "abstract".
        ?currPart   oa:hasBody      <http://pt.dbpedia.org/resource/Tomada_de_Decisão>.
        <http://pt.dbpedia.org/resource/Tomada_de_Decisão>  dbp-prop:label ?desc1
    }
}
```

Filter query results — Showing results from 1 to 7 of 7. Query took 1.29 s.

	currPart	desc1
1	wss_rdfa:6366851588892304?tags=formAcad	Engenharia Elétrica
2	wss_rdfa:4390865555343440?tags=formAcad	Engenharia Elétrica
3	wss_rdfa:6366851588892304?tags=formAcad	Banco de Dados
4	wss_rdfa:9408286832772730?tags=formAcad	Banco de Dados
5	wss_rdfa:6726713901235343?tags=formAcad	Banco de Dados
6	wss_rdfa:3564597309576489?tags=formAcad	Banco de Dados
7	wss_rdfa:9431313605945669?tags=abstract	Tomada de Decisão

Fonte:Próprio Autor.

A figura 58 demonstra o resultado e consulta da seguinte questão: Quais são as universidades que estão na Área de Formação Acadêmica dos currículos e quantos currículos tem formação nessa universidade?

Figura 58 – Consulta 04 no Sistema Lattes Web Semântico

```
PREFIX wss: <http://www.wssistemas.com.br/latteswss/>
PREFIX dbp-prop: <http://dbpedia.org/property/>
PREFIX schema: <http://schema.org/>
SELECT ?descUniv    ?o      (count(?s) as ?qntdCurr)
WHERE {
    ?s wss:pertence                 "formAcad".
    ?s wss:nomeInstituicaoEscola    ?o .
    ?o a                            schema:University.
    ?o dbp-prop:label               ?descUniv
        FILTER REGEX(str(?o),"pt.dbpedia")
}
Group By ?o ?descUniv ?numPosGrad
order By ?descUniv ?qntdCurr
```

Filter query results — Showing results from 1 to 26 of 26. Query took 0.277 s.

	descUniv	linkDbp	qntdCurr
1	Centro Federal de Educação Tecnológica Celso Suckow da Fonseca	http://pt.dbpedia.org/resource/Centro_Federal_de_Educação_Tecnológica_Celso_Suckow_da_Fonseca	"1"^^xsd:integer
2	Centro Federal de Educação Tecnológica de Minas Gerais	http://pt.dbpedia.org/resource/Centro_Federal_de_Educação_Tecnológica_de_Minas_Gerais	"1"^^xsd:integer
3	Centro Universitário do Espírito Santo	http://pt.dbpedia.org/resource/Centro_Universitário_do_Espírito_Santo	"1"^^xsd:integer
4	Pontifícia Universidade Católica de Minas Gerais	http://pt.dbpedia.org/resource/Pontifícia_Universidade_Católica_de_Minas_Gerais	"2"^^xsd:integer
5	Pontifícia Universidade Católica do Rio de Janeiro	http://pt.dbpedia.org/resource/Pontifícia_Universidade_Católica_do_Rio_de_Janeiro	"3"^^xsd:integer
6	San Diego State University	http://pt.dbpedia.org/resource/San_Diego_State_University	"1"^^xsd:integer
7	Universidade Candido Mendes	http://pt.dbpedia.org/resource/Universidade_Candido_Mendes	"1"^^xsd:integer
8	Universidade Estadual de Campinas	http://pt.dbpedia.org/resource/Universidade_Estadual_de_Campinas	"1"^^xsd:integer
9	Universidade Estácio de Sá	http://pt.dbpedia.org/resource/Universidade_Estácio_de_Sá	"1"^^xsd:integer
10	Universidade FUMEC	http://pt.dbpedia.org/resource/Universidade_FUMEC	"2"^^xsd:integer
11	Universidade Federal de Itajubá	http://pt.dbpedia.org/resource/Universidade_Federal_de_Itajubá	"2"^^xsd:integer
12	Universidade Federal de Minas Gerais	http://pt.dbpedia.org/resource/Universidade_Federal_de_Minas_Gerais	"7"^^xsd:integer

Fonte:Próprio Autor.

A próxima consulta é uma extensão da anterior, porém com um grande detalhe: a inserção da busca de uma informação que não existe nos currículos Lattes. A figura 59 exemplifica a utilização do LOD nos sistemas que a utilizam. Como temos os termos anotados em nosso sistema, podemos buscar outras informações em qualquer base aberta relacionada ao termo anotado. A figura 59 demonstra o resultado e consulta da seguinte questão: Mostrar a quantidade de pós-graduados dos currículos selecionados na consulta anterior, figura 58.

Figura 59 – Consulta 05 no Sistema Lattes Web Semântico

```
PREFIX wss: <http://www.wssistemas.com.br/latteswss/>
PREFIX dbp-prop: <http://dbpedia.org/property/>
PREFIX schema: <http://schema.org/>
PREFIX dbp-onto: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns>
SELECT ?descUniv    ?link       (count(?s) as ?qntdCurr)    ?numPosGrad
WHERE {
    ?s wss:pertence                 "formAcad".
    ?s wss:nomeInstituicaoEscola    ?link .
    ?link a                         schema:University.
    ?link dbp-prop:label                ?descUniv
        FILTER REGEX(str(?o),"pt.dbpedia")
    Service <http://pt.dbpedia.org/sparql>{

        #?o rdf:type        <http://dbpedia.org/ontology/University>;
        ?o  <http://dbpedia.org/ontology/numberOfPostgraduateStudents>  ?numPosGrad
    }
}
Group By ?o ?descUniv ?numPosGrad
order By ?descUniv ?qntdCurr
```

Filter query results — Showing results from 1 to 13 of 13. Query took 18.468 s.

	descUniv	link	qntdCurr	numPosGrad
1	Centro Federal de Educação Tecnológica Celso Suckow da Fonseca	http://pt.dbpedia.org/resource/Centro_Federal_de_Educação_Tecnológica_Celso_Suckow_da_Fonseca	"1"^^xsd:integer	"135"^^xsd:integer
2	Centro Federal de Educação Tecnológica de Minas Gerais	http://pt.dbpedia.org/resource/Centro_Federal_de_Educação_Tecnológica_de_Minas_Gerais	"1"^^xsd:integer	"970"^^xsd:integer
3	Pontifícia Universidade Católica do Rio de Janeiro	http://pt.dbpedia.org/resource/Pontifícia_Universidade_Católica_do_Rio_de_Janeiro	"3"^^xsd:integer	"7500"^^xsd:integer
4	Universidade Estadual de Campinas	http://pt.dbpedia.org/resource/Universidade_Estadual_de_Campinas	"1"^^xsd:integer	"26869"^^xsd:integer
5	Universidade Federal de Itajubá	http://pt.dbpedia.org/resource/Universidade_Federal_de_Itajubá	"2"^^xsd:integer	"374"^^xsd:integer
6	Universidade Federal de Minas Gerais	http://pt.dbpedia.org/resource/Universidade_Federal_de_Minas_Gerais	"7"^^xsd:integer	"14"^^xsd:integer
7	Universidade Federal de Minas Gerais	http://pt.dbpedia.org/resource/Universidade_Federal_de_Minas_Gerais	"7"^^xsd:integer	"838"^^xsd:integer
8	Universidade Federal de Santa Catarina	http://pt.dbpedia.org/resource/Universidade_Federal_de_Santa_Catarina	"2"^^xsd:integer	"8543"^^xsd:integer
9	Universidade Federal de Viçosa	http://pt.dbpedia.org/resource/Universidade_Federal_de_Viçosa	"2"^^xsd:integer	"2258"^^xsd:integer
10	Universidade Federal do Estado do Rio de Janeiro	http://pt.dbpedia.org/resource/Universidade_Federal_do_Estado_do_Rio_de_Janeiro	"1"^^xsd:integer	"1550"^^xsd:integer
11	Universidade Federal do Rio de Janeiro	http://pt.dbpedia.org/resource/Universidade_Federal_do_Rio_de_Janeiro	"1"^^xsd:integer	"9964"^^xsd:integer
12	Universidade de Brasília	http://pt.dbpedia.org/resource/Universidade_de_Brasília	"1"^^xsd:integer	"9905"^^xsd:integer

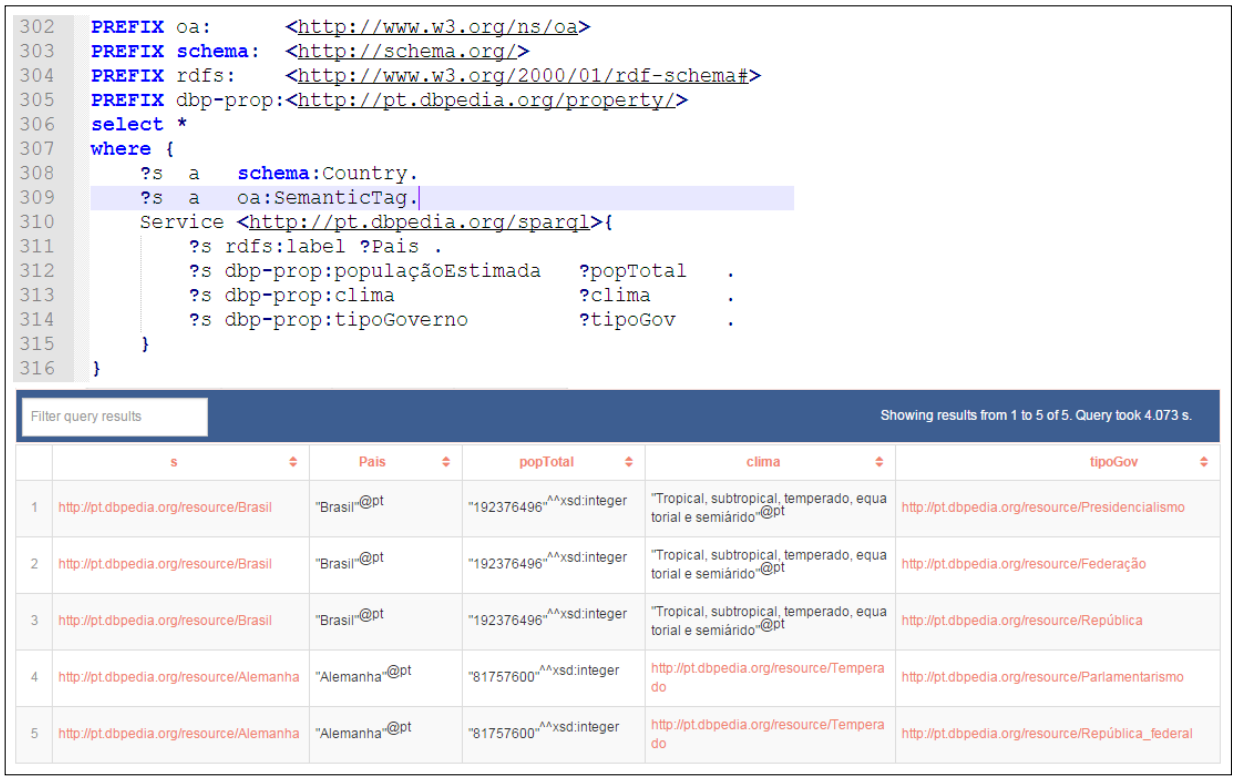
Fonte:Próprio Autor.

Uma observação a respeito do resultado da consulta demonstrada na figura 59 é com relação ao valor do dado retornado da página do dbpedia (pt). Existe a possibilidade da página está desatualizada com relação a informação desejada. Trata-se de uma questão de gerenciamento ou maneio das páginas que se encontram no LOD.

A última consulta visa abordar outro exemplo da correlação de informações que

estão nos currículos com informações externas do LOD. A figura 60 representa a resposta e a seguinte consulta: Quais são as anotações encontradas nos currículos Lattes que são classificados como um país e desses, represente a sua quantidade populacional, o clima e seu tipo de governo:

Figura 60 – Consulta 06 no Sistema Lattes Web Semântico



```
PREFIX oa:      <http://www.w3.org/ns/oa>
PREFIX schema:  <http://schema.org/>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbp-prop:<http://pt.dbpedia.org/property/>
select *
where {
    ?s  a   schema:Country.
    ?s  a   oa:SemanticTag.
    Service <http://pt.dbpedia.org/sparql>{
        ?s rdfs:label ?Pais .
        ?s dbp-prop:populaçãoEstimada   ?popTotal   .
        ?s dbp-prop:clima               ?clima      .
        ?s dbp-prop:tipoGoverno         ?tipoGov    .
    }
}
```

Filter query results — Showing results from 1 to 5 of 5. Query took 4.073 s.

	s	Pais	popTotal	clima	tipoGov
1	http://pt.dbpedia.org/resource/Brasil	"Brasil"@pt	"192376496"^^xsd:integer	"Tropical, subtropical, temperado, equatorial e semiárido"@pt	http://pt.dbpedia.org/resource/Presidencialismo
2	http://pt.dbpedia.org/resource/Brasil	"Brasil"@pt	"192376496"^^xsd:integer	"Tropical, subtropical, temperado, equatorial e semiárido"@pt	http://pt.dbpedia.org/resource/Federação
3	http://pt.dbpedia.org/resource/Brasil	"Brasil"@pt	"192376496"^^xsd:integer	"Tropical, subtropical, temperado, equatorial e semiárido"@pt	http://pt.dbpedia.org/resource/República
4	http://pt.dbpedia.org/resource/Alemanha	"Alemanha"@pt	"81757600"^^xsd:integer	http://pt.dbpedia.org/resource/Temperado	http://pt.dbpedia.org/resource/Parlamentarismo
5	http://pt.dbpedia.org/resource/Alemanha	"Alemanha"@pt	"81757600"^^xsd:integer	http://pt.dbpedia.org/resource/Temperado	http://pt.dbpedia.org/resource/República_federal

Fonte:Próprio Autor.

As consultas apresentadas visam demonstrar o potencial das anotações realizadas pelo Sistema Lattes Web Semântico com a utilização do LOD. A possibilidade de relacionarmos as fontes de dados abertas, Linked Open Data, com a base de dados desse sistema amplia a possibilidade de trabalharmos dados de domínios diferentes que não foram pesquisados.

8 Conclusão

A implantação da Web Semântica possibilita aumentar o significado de um conteúdo, tornando-o interpretável por humanos e aplicações. Como consequência, viabilizar um maior entendimento da estrutura do documento e recuperação de informações mais precisas. A maneira pela qual podemos implantar uma Web semanticamente compreensível por pessoas e computadores é por meio da anotação semântica. Um esquema para geração e uso de metadados possibilitando outros métodos de acesso a um dado ou informação.

Nessa pesquisa foi desenvolvido um Sistema, Lattes Web Semântico [1], que executa anotação automática para os documentos da Plataforma do Currículo Lattes do Cnpq utilizando dados de bases abertas, denominada de Linked Open Data. Afim de responder a questão do problema dessa pesquisa, podemos afirmar que para o desenvolvimento desse sistema os principais conceitos da Web Semântica que estão envolvidos são Extratores de Entidade, RDFa, RDF e SPARQL. O sistema é composto por um componente de extração de entidade, um anotador semântico e um componente de armazenamento e consulta de dados. Possibilita a disponibilização e indexação dos Currículos Lattes na Web através dos documentos HTML com RDFa embutido, disponibiliza as informações em RDF Turtle para serem utilizados em outros sistemas e domínios, por fim, expande a possibilidade de buscadores semânticos atuarem sobre esses currículos.

Criar um sistema de anotação semântica não foi uma tarefa trivial. É obrigatório devido a quantidade de documentos que há no cenário Lattes, a existência de ferramentas que trabalhem de forma automática (NETO, 2009). Na Revisão Sistemática da Literatura foi obtido a informação do Extrator de Entidade TextRazor que possibilita encontrar entidades em textos da língua português de forma automática. Na RSL encontrou-se as informações necessárias para a compreensão do cenário das bases abertas (LOD), seu funcionamento, acesso e particularidades; entendimento de uma outra estrutura de armazenamento de dados e consulta por meio de banco de dados de triplas utilizando a linguagem SPARQL.

Diante da experiência adquirida nessa pesquisa podemos refletir sobre algumas questões relacionadas ao desenvolvimento da Web Semântica no cenário brasileiro. Primeiro, é possível observar a necessidade de aumentar os grupos de trabalho que possam atuar em colaboração para alavancar o mapeamento do Wikipedia (pt.wikipedia.org) no LOD (pt.dbpedia.org) e gerenciar a manutenção do seu funcionamento. Necessidade da criação de um vocabulário formalizado pela instituição, CNPQ, para o Currículo Lattes,

[1] Disponível em www.wssistemas.com.br/latteswss/

permitindo uma efetiva inserção de seus dados nas bases abertas.

Com os estudos da RSL, desenvolvimento das etapas da concepção do sistema e observando os objetivos de cada componente, assim como apresentado por (REEVE; HAN, 2005), posso concluir que para uma plataforma de anotação semântica (PAS) o componente de extração de entidade (EE) é um objeto de grande relevância para o sucesso das anotações. Atentar que para o futuro das PAS, que terão como papel oferecer suporte, dados, para as máquinas de busca arquitetadas como buscadores semânticos, podemos dizer que anotar corretamente é disponibilizar informações integras, completas e corretas sobre um termo. Então, mais importante do que velocidade para encontrar entidades durante uma anotação automática de texto é a certeza de que o termo/palavra está caracterizada na entidade correta.

Diante das limitações desse projeto de pesquisa, podemos recomendar os seguintes trabalhos futuros:

1. Criação de um extrator de entidade específico para a língua portuguesa.
2. Utilização da junção de outros extratores de entidades para otimizar a quantidade de identificação dos termos/palavras nos LOD's.
3. Criar interface de operação de anotação manual para o usuário, possibilitando aumentar o número de termos anotados.
4. Continuar a extração das outras partes do Lattes, disponibilizando-as para a anotação semântica.
5. Verificar a possibilidade de interligar com os dados do projeto do Portal Brasileiro de Dados Abertos.
6. Utilizar agentes de pesquisa inteligente (buscador semântico) sob os documentos do lattes semântico.

O arcabouço desenvolvido nessa pesquisa demonstra o funcionamento da utilização do RDFa, RDF, SPARQL, Linked Open Data e sua efetividade para o alcance da Web Semântica. Atendeu o objetivo do LOD em identificar conjuntos de dados disponíveis sob licenças abertas (documentos Lattes) e convertê-los para RDF de acordo com os princípios Linked Data (BIZER TOM HEATH, 2009). O objetivo dessa pesquisa foi alcançado, pois os objetivos específicos descritos na seção 1.4.2 foram respondidos no capítulo 3, aplicados na seção 5 e no capítulo 6 de implementação.

Referências

1.1, W. S. Sparql 1.1 overview. In: . [s.n.], 2013. Disponível em: <https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/>.

BELLOZE, K. T. et al. An evaluation of annotation tools for biomedical texts. In: CITESEER. *ONTOBRAS-MOST*. 2012. p. 108–119. Disponível em: <http://ceur-ws.org/Vol-938/ontobras-most2012_paper9.pdf>.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web:a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, v. 284, n. 5, p. 34–43, may 2001. Disponível em: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>.

BIZER TOM HEATH, T. B. C. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, v. 5, n. 3, p. 1–22, 2009. Disponível em: <http://dx.doi.org/10.4018/jswis.2009081901>.

BONIFACIO, A. S. *Ontologias e consulta semântica : uma aplicação ao caso Lattes*. Dissertação (Mestrado) — UFRGS, Porto Alegre, 2002. Disponível em: <http://www.lume.ufrgs.br/handle/10183/7082>.

BRASIL, W. Web semântica. In: . [s.n.], 2014. Disponível em: <http://www.w3c.br/Padroes/WebSemantica>.

BUTUC, M.-G. Semantically enriching content using opencalais. In: . [s.n.], 2009. Disponível em: <http://www.eed.usv.ro/SistemeDistribuite/2009/Butuc1.pdf?origin=publication_detail>.

CASARIN, H. d. C. S.; CASARIN, S. J. C. *Pesquisa Científica: da teoria à prática*. [S.l.: s.n.], 2011.

CASTAñO, A. C. *Populando ontologias através de informações em HTML - o caso do currículo lattes*. Dissertação (Mestrado) — Universidade de São Paulo, 2008. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45134/tde-12082008-130204/>.

CERVO, A.; BERVIAN, P.; SILVA, R. da. *Metodologia científica*. Pearson Prentice Hall, 2006. ISBN 9788576050476. Disponível em: <https://books.google.com.br/books?id=9SK2GQAACAAJ>.

CNPQ. Plataforma lattes cnpq. In: . [s.n.], 2014. Disponível em: <http://lattes.cnpq.br/>.

DERCZYNSKI DIANA MAYNARD, G. R. M. v. E. G. G. R. T. J. P. K. B. L. Analysis of named entity recognition and linking for tweets. *Information Processing and Management: www.elsevier.com/locate/infoproman*, n. 17, p. 32–49, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457314001034>.

ELLER, M. P. *Anotações Semânticas de Fontes de Dados Heterogêneas Um Estudo de Caso com a Ferramenta Smore*. Dissertação (Mestrado) — Universidade Federal

de Santa Catarian – Departamento de Informática e Estatística, 2008. Disponível em: <https://projetos.inf.ufsc.br/arquivos_projetos/projeto_752/TCC_Markus.pdf>.

FAFALIOS, P.; PAPADAKOS, P. Theophrastus: On demand and real-time automatic annotation and exploration of (web) documents using open linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, n. 0, p. –, 2014. ISSN 1570-8268. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1570826814000572>.

FONTES, C. A.; CAVALCANTI, M.; MOURA, A. D. C. An ontology based reasoning approach for document annotation. p. 160–167, Sept 2013. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6693512>.

FONTES, C. A.; MOURA, A. M. de C.; CAVALCANTI, M. C. Anotacao semantica em documentos. In: . [s.n.], 2010. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/wtdbd/2010/sbbd_wtd_14.pdf>.

FONTES, C. A. et al. Recuperação de informações em documentos anotados semanticamente na Área de gestão ambiental. p. 43–52, 2010. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/ontobras/2010/004.pdf>.

GALEGO, E. F. Extração e consulta de informações do currículo lattes baseada em ontologias. 2013. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0043.pdf>.

GIL, A. C. *Como elaborar projetos de pesquisa*. [S.l.: s.n.], 2002.

GOV, W. D. A. Dados abertos governamentais. In: . [s.n.], 2014. Disponível em: <http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf>.

GROUP, W. O. A. C. Open annotation data model. In: . [s.n.], 2013. Disponível em: <http://www.openannotation.org/,http://www.commonsemantics.com/oa/Open%20Annotation%20Data%20Model%20Primer.html,http://www.openannotation.org/spec/core/20130208/index.html>.

MASCARENHAS, S. *METODOLOGIA CIENTIFICA*. PEARSON BRASIL. ISBN 9788564574595. Disponível em: <https://books.google.com.br/books?id=kOZBLgEACAAJ>.

MENDES, P. N. et al. Dbpedia spotlight: Shedding light on the web of documents. ACM, New York, NY, USA, p. 1–8, 2011. Disponível em: <http://doi.acm.org/10.1145/2063518.2063519>.

MUNARO, B.; LIMA, M. L.; CAMPOS, M. Recomendação de dados abertos para solucionar os problemas de comunicação textual : uma análise de métodos para extração de entidades nomeadas. 2012. Disponível em: <http://www.imago.ufpr.br/csbc2012/anais_csbc/eventos/brasnam/artigos/BRASNAM%20-%20Recomendacao%20de%20dados%20abertos%20para%20solucionar%20os%20problemas%20de%20comunica%C3%A7%C3%A3o%20textual%20uma%20analise%20de%20metodos%20para%20extracao%20de%20entidades%20nomeadas.pdf>.

NETO, G. M. d. S. *Anotacao Semantica De Recursos Web Baseada em Ontologias*. Dissertação (Mestrado) — Dissertação de Mestrado–UFAM–INSTITUTO DE CIÊNCIAS EXATAS–PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA, 2009. Disponível em: <http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=148061>.

ONTOTEXT. Graphdb ontotext. In: . [s.n.], 2015. Disponível em: <http://ontotext.com/products/graphdb/>.

OREN, E. et al. What are semantic annotations. In: . [S.l.]: Citeseer, 2006.

OWL1, W. Owl web ontology language. In: . [s.n.], 2014. Disponível em: <http://www.w3.org/TR/owl-features/>.

OWL2, W. Owl2 - web ontology language 2. In: . [s.n.], 2015. Disponível em: <http://www.w3.org/TR/2012/REC-owl2-overview-20121211/>.

PLANETDATA. Linked open data cloud diagram 2014. In: . [s.n.], 2014. Disponível em: <http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/>.

PRIMER, W. R. Rdfa 1.1 primer - second edition. In: . [s.n.], 2014. Disponível em: <http://www.w3.org/TR/2013/NOTE-rdfa-primer-20130822/>.

REEVE, L.; HAN, H. Survey of semantic annotation platforms. In: *Proceedings of the 2005 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2005. (SAC '05), p. 1634–1638. ISBN 1-58113-964-0. Disponível em: <http://doi.acm.org/10.1145/1066677.1067049>.

SALEH, L. M. B.; AL-KHALIFA, H. S. Aratation: An arabic semantic annotation tool. ACM, New York, NY, USA, p. 447–451, 2009. Disponível em: <http://doi.acm.org/10.1145/1806338.1806421>.

SCHEMA1.1, W. Rdf schema 1.1. In: . [s.n.], 2014. Disponível em: <http://www.w3.org/TR/2014/REC-rdf-schema-20140225/>.

TAO, C. et al. Semantator: Semantic annotator for converting biomedical text to linked data. *Journal of Biomedical Informatics*, v. 46, n. 5, p. 882–893, 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1532046413001020>.

TEXTRAZOR. Textrazor. In: . [s.n.], 2015. Disponível em: <https://www.textrazor.com/>.

VIRGILIO, R. D. et al. A reverse engineering approach for automatic annotation of web pages. *Multimedia Tools and Applications*, v. 64, n. 1, p. 119–140, may 2013. ISSN 1380-7501, 1573-7721. 00001. Disponível em: <http://link.springer.com/article/10.1007/s11042-011-0852-8>.

W3C-RDF1.1-PRIMER. Rdf 1.1 primer. In: . [s.n.], 2014. Disponível em: <http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/>.

WEB, C. de Estudos sobre T. Centro web brasil. In: . [s.n.], 2016. Disponível em: <http://ceweb.br/>.

ZHANG, Z.; CHEN, S.; FENG, Z. Semantic annotation for web services based on DBpedia. p. 280–285, 2013. Disponível em: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=6525532>.

Apêndices

APÊNDICE A – Produções Selecionadas na RSL

Tabela 4 – Principais Produções Selecionadas Para Leitura Completa na Revisão Sistemática da Literatura

Título Produção	**Objetivo com Trabalho**
Theophrastus: On demand and real-time automatic annotation and exploration of (web) documents using open linked data	5
Semantator: Semantic annotator for converting biomedical text to linked data	5
Semantic Annotation Framework For Intelligent Information Retrieval Using KIM	5
Semantically Enriching Content Using OpenCalais	5
DBpedia spotlight- shedding light on the web of documents	5
A machine Learning Based Analytical Framework For Semantic Annotation Requirmnts 2011	5
Semantic annotation tools survey	5
Automated Semantic Tagging of Textual Content	5
Survey of Semantic Annotation Platforms	5
Anotacao Semantica De Recursos Web Baseada em Ontologias-UFAM	5
An Evaluation of Annotation Tools for Biomedical Texts	5
Semantic Annotation for Web Services Based on Dbpedia	5
Integrating Keywords and Semantics on Document Annotation and Search	4
GonTogle: A Tool for Semantic Annotation and Search	4
BioAnnote: A software platform for annotating biomedical documents with application in medical learning environments	4
A Proposed Framework for Arabic Semantic Annotation Tool	4
NCBO Annotator: Semantic Annotation of Biomedical Data	4
Recuperação de Informacoes em Documentos Anotados Semanticamente na GesAmb	4
Analysis of named entity recognition and linking for tweets	4
AraTation: an Arabic semantic annotation tool	4

Fonte: Próprio Autor.

APÊNDICE B – Código Fonte:Mapeamento RDFa

```
1        <?php
2        function rdfa_cabecalho($idLattes, $nomeAutorLattes,
            $siteLattesCnpq){
3             $cabecalho='<!DOCTYPE html PUBLIC "-//W3C//DTD
                 XHTML+RDFa 1.1//EN" "http://www.w3.org/MarkUp/
                 DTD/xhtml-rdfa-2.dtd">
4        <html version="XHTML5+RDFa 1.1" xml:lang="pt" lang="pt"
            xmlns="http://www.w3.org/1999/xhtml">
5        <head>
6            <meta http-equiv="Content-Type" content="application/
                xhtml+xml; charset=utf-8" />
7            <meta name="content-language" content="pt" />
8            <meta content="Sistema Lattes Web Sem ntico" http-
                equiv="generator"/>
9            <link rel="stylesheet" href="../css/anotacao.css"
                type="text/css"/>
10           <title>Curr culo Sem ntico Anotado em RDFa do
                Sistema de Curr culos Lattes ('.$nomeAutorLattes.'
                )</title>
11       </head>
12       <body prefix=
13               "cv: http://rdfs.org/resume-rdf/cv.rdfs#
14               wss: http://www.wssistemas.com.br/latteswss/
15               wss_rdfa: http://www.wssistemas.com.br/latteswss/
                    recursordfa/
16               lattes: '.$siteLattesCnpq.'
17               dc: http://purl.org/dc/terms/
18               foaf: http://xmlns.com/foaf/0.1/
19               schema: http://schema.org/
20               rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
21               dbp-onto: http://dbpedia.org/ontology/
22               dbp-prop: http://dbpedia.org/property/ " about="
                    wss_rdfa:'.$idLattes.'" typeof="cv:CV">
23
24               ';
```

```
25              return $cabecalho;
26      }
27
28      function rdfa_AutorLattes($idLattes, $nomeAutorLattes){
29          $resultado='
30              <div class="rdfa_autorLattes">
31                      <span property="cv:cvTitle">Curr culo
                           Lattes Anotado em RDFa do Sistema de
                           Curr culos Lattes ('.$nomeAutorLattes.
                           ')</span>
32                      N mero Identificador:
33                      <span rel="foaf:givenName">'. $idLattes.'
                           </span>
34                      Autor Lattes:
35                      <span property="wss:AutorLattes">'.
                           $nomeAutorLattes.'</span>
36                      <span rel="cv:aboutPerson dc:creator"
                           typeof="schema:Person dbp-onto:Person
                           foaf:Person" resource="wss:'.
                           $nomeAutorLattes.'">'.$nomeAutorLattes.
                           '</span>
37                      <span rel="schema:sameAs" resource="
                           lattes:'.$idLattes.'" typeof="cv:CV"
                           ></span>
38                      <span property="dc:format" resource="html
                           /rdfa" ></span>
39              </div>
40              <BR>
41              ';
42
43          return $resultado;
44      }
45
46      function rdfa_resumoCurr($idCurrLattes, $txtResumoCurr){
47
48              $resultado = '
49              <div class="abstract" resource="wss_rdfa:'.
                   $idCurrLattes.'?tags=abstract" >
50                      <span rel="wss:eParte"  resource:"
                           wss_rdfa:'.$idCurrLattes.'"></span>
51                      <span property="wss:pertence" about="
                           abstract"></span>
```

```
52                  ';
53
54                  $resultado.= $txtResumoCurr;
55                  $resultado.= '
56                  </div>
57                  <BR><BR>
58                  ';
59
60                  return $resultado;
61
62          }
63
64          function rdfa_Year($ano){
65              $resultado='
66              <div class="rdfa_Year" typeof="rdf: http://dbpedia.
                    org/ontology/Year">
67                  <span property="rdf:label">'.$ano.'</span>';
68              $resultado.="
69              </div> \n";
70
71              return $resultado;
72
73          }
74
75          function rdfa_resourceFreebase($FreeBase_ID){
76                  if(!empty($FreeBase_ID))
77                          $resultado = '<link property="schema:
                                sameAs" href="http://www.freebase.com/'
                                .$FreeBase_ID.'?lang=pt"/> /n';
78                  else
79                          echo "Fun   o rdfa_resourceFreebase:
                                Vari vel FreeBase_ID vazia (
                                $FreeBase_ID)!";
80                  return $resultado;
81          }
82
83          function rdfa_Organization($nome,$resource,$FreebaseID)
84          {
85              $resultado='
86              <div class="rdfa_Organization" typeof="schema:
                    Organization">
87                  <a property="url" href="'.$resource.'">
```

```
88                      <span property="schema:name">'.$nome.'</
                           span>
89                 </a>
90                 <link property="schema:sameAs" href="'.$resource.'
                      "/>';
91             if(!empty($FreeBase_ID)) $resultado .=
                   rdfa_resourceFreebase($FreebaseID);
92             $resultado.="
93             </div> \n";
94
95             return $resultado;
96         }
97         function rdfa_Company($nome,$resource,$FreebaseID)
98         {
99             $resultado='
100            <div class="rdfa_Company" resource="'.$resource.'"
                   typeof="schema:Company">
101                <a property="url" href="'.$resource.'">
102                   <span property="schema::name">'.$nome.'</
                         span>
103                </a>
104                <link property="schema:sameAs" href="'.$resource.'
                      "/>';
105            if(!empty($FreeBase_ID)) $resultado .=
                   rdfa_resourceFreebase($FreebaseID);
106
107            $resultado.="
108            </div> \n";
109
110            return $resultado;
111        }
112        function rdfa_pessoa($nome,$resource)
113        {
114            $resultado='
115            <div class="rdfa_pessoa" typeof="foaf:Person schema:
                   Person">
116                <a property="url" href="'.$resource.'">
117                       <span property="foaf:name">'.$nome.'</span>
118                </a>';
119
120            $resultado.="
121            </div> \n";
```

```
122
123             return $resultado;
124         }
125         function rdfa_ColegioUniversidade($colegUniversidade,
                $resource){
126
127                 $resultado='
128                 <div class="rdfa_ColegioUniversidade" about="'.
                        $resource.'" typeof="schema:CollegeOrUniversity
                        ">
129                         <a property="url" href="'.$resource.'">
130                                 <span property="schema:name">'.
                                        $colegUniversidade.'</span>
131                 </a>
132                 <link property="schema:sameAs" href="'.$resource.
                        '"/>
133             </div>';
134
135                 return $resultado;
136         }
137         function rdfa_Pais($pais,$resource){
138                 //http://schema.org/Country
139                 $resultado="\n".'
140                 <div class="rdfa_Pais" typeof="schema:Country">
141                 <a property="url" href="'.$resource.'">
142                     <span property="schema:name">'.$pais.'</span>
143                 </a>
144                 <link property="schema:sameAs" href="'.
                        $resource.'"/>';
145             $resultado.="
146             </div> \n";
147                 return $resultado;
148         }
149         function rdfa_Coisa($nomeCoisa,$resource,$FreebaseID){
150             $resultado="\n".'
151             <div class="rdfa_Coisa" typeof="schema:Thing">
152                 <a property="url" href="'.$resource.'">
153                     <span property="schema:name">'.$nomeCoisa.'</
                            span>
154                 </a>
155                 <link property="schema:sameAs" href="'.$resource.
                        '"/>';
```

```
156             if(!empty($FreeBase_ID)) $resultado .=
                    rdfa_resourceFreebase($FreebaseID);
157
158             $resultado.="
159             </div> \n";
160
161             return $resultado;
162         }
163
164         function rdfa_anotado($textoAnterior,$textoDepoisAnotado)
                //verifica se o texto foi anotado.
165         {
166             if( (md5($textoAnterior)!==md5($textoDepoisAnotado))
                    ){
167                 return true;
168             }else{
169                 return false;
170
171             }
172         }
```

APÊNDICE C – Código Fonte:Mapeamento RDF Turtle

```
1   <?php
2   function rdf_Prefixos(){
3
4           $resultado="@Prefix rdf:                          <
              http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
5   @Prefix rdfs:                   <http://www.w3.org
      /2000/01/rdf-schema#>.
6   @Prefix dbp-onto:               <http://dbpedia.org/
      ontology/>.
7   @Prefix dbp-prop:               <http://dbpedia.org/
      property/>.
8   @Prefix dc:                             <http://purl.org/
      dc/terms/>.
9   @Prefix schema:                 <http://schema.org/>.
10  @PREFIX skos:                   <http://www.w3.org
      /2004/02/skos/core#>.
11  @Prefix foaf:                   <http://xmlns.com/foaf
      /0.1/>.
12  @Prefix cv:                             <http://rdfs.org/
      resume-rdf/cv.rdfs#>.
13  @Prefix oa:                             <http://www.w3.
      org/ns/oa#>.
14  @Prefix cnt:                    <http://www.w3.org/2011/
      content#>.
15  @Prefix prov:                   <http://www.w3.org/ns/
      prov#>.
16  @Prefix lattes:                 <http://lattes.cnpq.br/>.
17  @Prefix wss:                    <http://www.wssistemas.
      com.br/latteswss/>.
18  @Prefix wss_rdfa:               <http://www.wssistemas.
      com.br/latteswss/recursordfa/>.
19
20  ################## Propriedades WSS
      ############################[InsertTriplas]
21  wss:autorLattes         rdf:type        rdfs:Property.
```

```
22          wss:autorLattes          rdfs:domain     dbp-onto:Person.
23          wss:eParte                       rdf:type        rdfs:
               Property.
24          wss:pertence             rdf:type        rdfs:Property.
25          wss:temAnotado           rdf:type        rdfs:Property.
26          wss:temAnotado           rdf:domain      oa:Annotation, oa
               :d4e434.
27          wss:temAnotado           rdf:range       oa:Tag, oa:d4e654
               , cnt:ContentAsText.
28          ###############Propriedades Area Formacao Academica
29                          wss:grauAcad
                               rdf:type                skos:Concept.
30                          wss:grauAcad
                               dc:subject      <http://pt.dbpedia.org/
                               page/Categoria:Graus_acad micos>.
31                          ";
32          foreach (rdf_PrefixosRdf("formAcad") as $campo){
33                          $resultado .= "wss:$campo
                                              rdf:type
                                        rdfs:Property.
34                          ";
35          }
36
37          $resultado .= "################## Fim Propriedades WSS
               ############################
38
39
40          ";
41
42
43                  return $resultado;
44          }
45          function rdf_PrefixosRdf($areaCurrLattes){
46
47                  if($areaCurrLattes=="formAcad"){
48                          $arrayCampos = array(
49                                          "nomeOrientador",
50                                          "nomeCoOrientador",
51                                          "nomeInstituicaoEscola",
52                                          "
                                               nomeInstituicaoEscolaOutra
                                               ",
```

```
53                                      "nomeCurso",
54                                      "nomeCursoEN",
55                                      "anoInicio",
56                                      "anoFim",
57                                      "tituloProducao",
58                                      "tituloProducaoEN",
59                                      "nomeCursoEN"   ,
60                                      "
                                           nomeInstituicaoEscolaGrad
                                           ",
61                                      "nomeOrientadorGrad",
62                                      "palavrasChave"
63                      );
64              }
65
66              return $arrayCampos;
67
68      }
69      function rdf_addTipoAnotacao($tipoEntidade){ //Retorna o(
           s) tipo de entidade(s)
70
71              $tpEntidadeWiki=null;
72              if($tipoEntidade!="ND" and $tipoEntidade!=""){
73                      if($tipoEntidade=="Person"){
74                              $tpEntidadeWiki = ", schema:
                                   Person, foaf:Person, dbp-onto:
                                   Person";
75                      }elseif($tipoEntidade==""
                           CollegeOrUniversity" or $tipoEntidade==
                           "University" or $tipoEntidade==""
                           EducationalInstitution"){
76                              $tpEntidadeWiki = ", schema:
                                   CollegeOrUniversity, schema:
                                   University, schema:
                                   EducationalInstitution";
77                      }elseif($tipoEntidade=="Company"){
78                              $tpEntidadeWiki = ", schema:
                                   Company, dbp-onto:Company, foaf
                                   :Company";
79                      }elseif($tipoEntidade=="Country"){
80                              $tpEntidadeWiki = ", schema:
                                   Country, dbp-onto:Country";
```

```
81                          }elseif($tipoEntidade=="City"){
82                                  $tpEntidadeWiki = ", schema:City,
                                        dbp-onto:City";
83                          }elseif($tipoEntidade=="State"){
84                                  $tpEntidadeWiki = ", schema:State
                                        , dbp-onto:State";
85                          }elseif($tipoEntidade=="Continent"){
86                                  $tpEntidadeWiki = ", schema:
                                        Continent, dbp-onto:Continent";
87                          }elseif($tipoEntidade=="Organisation" or
                                $tipoEntidade=="Organization"){
88                                  $tpEntidadeWiki = ", schema:
                                        Organisation, dbp-onto:
                                        Organisation, foaf:Organization
                                        ";
89                          }elseif($tipoEntidade=="Year"){
90                                  $tpEntidadeWiki = ", schema:Year"
                                        ;
91                          }elseif($tipoEntidade=="Place"){
92                                  $tpEntidadeWiki = ", schema:Place
                                        , dbp-onto:Place";
93                          }else{
94                                  $tpEntidadeWiki = ", schema:".
                                        $tipoEntidade;
95                                  echo "Fun   o
                                        rdf_AddTipoAnotacao: Tipo da
                                        Entidade definida
                                        autom ticamente ($tipoEntidade
                                        )";
96                          }
97
98                  }Else
99                          $tpEntidadeWiki= "";
100
101                 return $tpEntidadeWiki;
102         }
103
104         function rdf_addPropriedadesSemanticas($tipoEntidade,
                $vlEntidade){
105
106                 $novaPropriedadeSemantica=null;
107                 if($tipoEntidade!="ND" and $tipoEntidade!=""){
```

```
108                         if($tipoEntidade=="Person"){
109                                 $novaPropriedadeSemantica =
110                                 "\n             dbp-prop:nome
                                                    \"$vlEntidade
                                       \";
111                         dbp-prop:nomeCompleto   \"$vlEntidade\";
112                         foaf:name
                               \"$vlEntidade\"; ";
113                         }elseif($tipoEntidade==""
                               CollegeOrUniversity" or $tipoEntidade==
                               "University" or $tipoEntidade==""
                               EducationalInstitution"){
114                                 $novaPropriedadeSemantica= "";
115                         }elseif($tipoEntidade=="Company"){
116                                 $novaPropriedadeSemantica = "";
117                         }elseif($tipoEntidade=="Country"){
118                                 $novaPropriedadeSemantica = "";
119                         }elseif($tipoEntidade=="Continent"){
120                                 $novaPropriedadeSemantica = "";
121                         }elseif($tipoEntidade=="Organisation"){
122                                 $novaPropriedadeSemantica = "";
123                         }elseif($tipoEntidade=="Year"){
124                                 $novaPropriedadeSemantica = "";
125                         }/*else{
126                                 echo "Fun   o
                                       addPropriedadesSemanticas:
                                       Propriedades para a entidade  (
                                       $tipoEntidade) n  o definida.";
127                         }*/
128
129                 }else{
130                         $novaPropriedadeSemantica= "";
131                 }
132
133                 return $novaPropriedadeSemantica;
134         }
135
136
137         function rdf_linkWikiPT($linkWiki){ //entra um
               linkWikipedia EN ou PT e retorna em linkWikipediaPT
138
139                 $haystack = $linkWiki;
```

```
140                 $needle    = 'en.';
141
142                 $pos       = strripos($haystack, $needle);
143
144                 if ($pos === false) {
145                         //echo "Sinto muito, n s  n  o
                               encontramos ($needle) em ($haystack)";
146                         $linkWikiPT     = $linkWiki;
147                 } else {
148                         //echo "N  s  encontramos a  ltima  (
                               $needle) em ($haystack) na posi  o (
                               $pos)";
149                         $linkWikiPT     =str_replace("/en.
                               wikipedia", "/pt.wikipedia", $linkWiki)
                               ; //wiki em portugues
150                 }
151
152                 return $linkWikiPT;
153
154         }
155
156         function rdf_linkWikiEN($palavra_Termo_AnotadoEN_){ //
               entra uma palavra/termo e retorna em linkWikipediaEN
157         /*
158                 $haystack = $linkWiki;
159                 $needle    = 'pt.';
160
161                 $pos       = strripos($haystack, $needle);
162
163                 if ($pos === false) {
164                         //echo "Sinto muito, n s  n  o
                               encontramos ($needle) em ($haystack)";
165                         $linkWikiEN     = $linkWiki;
166                 } else {
167                         $linkWikiEN     =str_replace("/pt.
                               wikipedia", "/en.wikipedia", $linkWiki)
                               ; //wiki em ingles
168                 }
169                 */
170                 $linkWikipediaEN        = "https://en.wikipedia.
                    org/wiki/".$palavra_Termo_AnotadoEN_;
171                 return $linkWikipediaEN;
```

```
172
173         }
174
175         function rdf_linkDBpediaPT($palavra_Termo_Anotado_){ //
               Entra termo/palavra em portugues ou ingles e retorna
               DBpediaPt.
176         /*
177                 $haystack = $linkWikipedia;
178                 $needle   = 'en.';
179
180                 $pos      = strripos($haystack, $needle);
181
182                 if ($pos === false) {
183                         //echo "Sinto muito, n s  n  o
                           encontramos ($needle) em ($haystack)";
184                         $linkDBpediaPT  =str_replace("/pt.
                           wikipedia.org/wiki/", "/pt.dbpedia.org/
                           resource/", $linkWikipedia); //DBpedia
                           em portugues
185                 } else {
186                         //echo "N  s  encontramos a  ltima  (
                           $needle) em ($haystack) na posi  o (
                           $pos)";
187                         $linkDBpediaPT  =str_replace("/en.
                           wikipedia.org/wiki/", "/pt.dbpedia.org/
                           resource/", $linkWikipedia); //DBpedia
                           em portugues
188                 }
189         */
190                 $palavra_Termo_Anotado_ = ucfirst(
                    $palavra_Termo_Anotado_);//primeira letra em
                    maiusculo.
191                 $linkDBpediaPT  = "http://pt.dbpedia.org/resource
                    /".$palavra_Termo_Anotado_;
192                 return $linkDBpediaPT;
193
194         }
195
196         function rdf_linkDBpediaEN($palavra_Termo_AnotadoEN_){ //
               Entra termo/palavra em ingles e retorna DBpediaEN.
197         /*
198                 $haystack = $linkWikipedia;
```

```
199                 $needle    = 'en.';
200
201                 $pos       = strripos($haystack, $needle);
202
203                 if ($pos === false) {
204                         //echo "Sinto muito, n s  n  o
                               encontramos ($needle) em ($haystack)";
205                         $linkDBpediaEN  =str_replace("/pt.
                               wikipedia.org/wiki/", "/dbpedia.org/
                               resource/", $linkWikipedia); //DBpedia
                               em portugues
206                 } else {
207                         //echo "N  s  encontramos a  ltima  (
                               $needle) em ($haystack) na posi  o (
                               $pos)";
208                         $linkDBpediaEN  =str_replace("/en.
                               wikipedia.org/wiki/", "/pt.dbpedia.org/
                               resource/", $linkWikipedia); //DBpedia
                               em portugues
209                 }
210         */
211                 $palavra_Termo_AnotadoEN_ = ucfirst(
                       $palavra_Termo_AnotadoEN_);//primeira letra em
                       maiusculo.
212                 $linkDBpediaEN  = "http://dbpedia.org/resource/".
                       $palavra_Termo_AnotadoEN_;
213                 return $linkDBpediaEN;
214
215         }
216
217         function rdf_AutorLattesTTL($numCVLattes,$nomeAutorLattes
                   ){
218                 $resultado=
219                 "
220
221         ################## Iniciando Cabe alho
222         <wss_rdfa:$numCVLattes>
223                         a                                   cv:CV;
224                         wss:AutorLattes \"$nomeAutorLattes\";
225                         cv:aboutPerson  <wss:".str_replace(" ", "
                               _", $nomeAutorLattes).">;
226                         cv:cvTitle              \"Curr culo
```

```
                        Lattes Anotado em RDF do Sistema de
                        Curr culos Lattes ($nomeAutorLattes)
                        \";
227                     dc:creator              \"
                        $nomeAutorLattes\";
228                     schema:sameAs   <lattes:$numCVLattes>;
229                     dc:format       \"text/html\".
230
231         <wss:".str_replace(" ", "_", $nomeAutorLattes).">
                    a               schema:Person, foaf:Person, dbp
            -onto:Person .
232
233                 ";
234                 return $resultado;
235         }
236
237         function rdf_CorpoTagSemanticTag($numCVLattes,
            $ondeLocalCV,$anotadoPorIndexSiteWSS,
            $anotadoPorIndexPessoaSite,$arrayEntdResumo){
238
239                 $IniciandoHasbody=
240                 "\n\n\n\n####################Iniciando o HasBody
                    do $ondeLocalCV:
241         <wss_rdfa:$numCVLattes?tags=".$ondeLocalCV.">   wss:
            eParte              <wss_rdfa:$numCVLattes>;
242
243
244
245         <wss_rdfa:$numCVLattes?tags=".$ondeLocalCV.">   a
            oa:Annotation, oa:d4e434 ;
246                     oa:hasTarget    <wss:/recursoxml/
                        $numCVLattes?tags=".$ondeLocalCV.">,<
                        lattes:$numCVLattes>;
247                     oa:annotatedBy  <$anotadoPorIndexSiteWSS
```

```
                        >,<$anotadoPorIndexPessoaSite>;
248                     oa:annotatedAt  \"".date("Y-m-d")."T".
                        date("H:i:s")."Z"."\" ;
249                     oa:serializedBy <$anotadoPorIndexSiteWSS
                        >,<https://www.textrazor.com/>;
250                     oa:serializedAt \"".date("Y-m-d")."T".
                        date("H:i:s")."Z"."\" ;";
251
252             //escrevendo o hasBody:
253             $hasbody = "\n\n#####HasBody do $ondeLocalCV: " .
                    rdf_hasBodyTag($numCVLattes, $ondeLocalCV,
                    $arrayEntdResumo);
254
255
256             //descrevendo cada tag do hasBody:
257             $desHasBody = "\n\n####Descri  o do HasBody do
                    $ondeLocalCV: " . rdf_DescHasBody($numCVLattes,
                    $ondeLocalCV, $arrayEntdResumo);
258
259             $temAnotadoToWss = rdf_AnotacoesWss($numCVLattes,
                    $ondeLocalCV,$arrayEntdResumo);
260
261             return $IniciandoHasbody. $hasbody .  $desHasBody
                    .$temAnotadoToWss."\n";
262     }
263
264     function rdf_hasBodyTag($numCVLattes,$ondeLocalCV,
            $arrayEntdResumo){
265
266             //escrevendo o hasBody:
267             $resultado="";
268             foreach ($arrayEntdResumo as $valores)//pegando
                    os valores, entidades, termos do resumo
269             {
270                     $tipoEntidadeWiki               = trim(
                        $valores['tipoEntidade']);
271                     $palavraTermoAnotado    = trim($valores['
                        valorEntidade']);
272                     $palavraTermoAnotadoEN  = trim($valores['
                        valorEntidadeEN']);
273                     $WikipediaURL   = trim($valores['
                        WikipediaURL']);
```

```
274                         $FreebaseID             = trim($valores['
                               FreebaseID']);
275
276                         $txtPalavra_Anotada_ = str_replace(" ", "
                               _", $palavraTermoAnotado);
277                         $txtPalavra_AnotadaEN_ = str_replace(" ",
                                "_", $palavraTermoAnotadoEN);
278
279                         $resultado.=
280                         "
281                         oa:hasBody <wss_rdfa:$numCVLattes?tags=
                               $ondeLocalCV?tag=$txtPalavra_Anotada_>;
                               ";
282
283                         //se existe a informa  o e    diferente
                                do r tulo wikipedia Url:
284                         if(strlen($WikipediaURL)!=0 and
                               $WikipediaURL!="Wikipedia Url"){
285                               $resultado              .="\n
                                                      oa:hasBody <".
                                  rdf_linkDBpediaPT(
                                  $txtPalavra_Anotada_).">;" ;
286                         }
287
288                         //se existe a informa  o e    diferente
                                do r tulo wikipedia Url:
289                         if(strlen($txtPalavra_AnotadaEN_)!=0 and
                               $txtPalavra_AnotadaEN_!="
                               valorEntidadeEN"){
290                               $resultado              .="\n
                                                      oa:hasBody <".
                                  rdf_linkDBpediaEN(
                                  $txtPalavra_AnotadaEN_).">;" ;
291                         }
292
293                         if(strlen($FreebaseID)!=0){
294                               $FreebaseID             = "http
                                  ://www.freebase.com/".
                                  $FreebaseID."?lang=pt";
295                               $resultado              .="\n
                                                      oa:hasBody <
                                  $FreebaseID>;";
```

```
296                         }
297
298                 }
299
300                 //Tirando a virgula da  ltima  linha e
                       adicionando o ponto final.
301                 $resultado .=".";
302                 $resultado = str_replace(">;.", ">.", $resultado)
                       ;
303
304                 return $resultado;
305         }
306
307         function rdf_DescHasBody($numCVLattes,$ondeLocalCV,
               $arrayEntdResumo){
308
309                 $resultado=null;
310                 //descrevendo o hasBody:
311                 foreach ($arrayEntdResumo as $valores)//pegando
                       os valores, entidades, termos do resumo
312                 {
313                         $tipoEntidadeWiki               = trim(
                               $valores['tipoEntidade']); //pode vir
                               vazio ou "ND" n o definido
314                         $palavraTermoAnotado    = trim($valores['
                               valorEntidade']); //sempre tem
                               conte do
315                         $palavraTermoAnotadoEN  = trim($valores['
                               valorEntidadeEN']); //sempre tem
                               conte do
316                         $WikipediaURL   = trim($valores['
                               WikipediaURL']); //pode vir vazio
317                         $FreebaseID             = trim($valores['
                               FreebaseID']); //pode vir vazio
318
319                         if($tipoEntidadeWiki=="tipoEntidade")
                               continue;
320
321                         //colocando _ nas palavras com espa o
322                         $txtPalavra_Anotada_ = str_replace(" ", "
                               _", $palavraTermoAnotado);
323                         $txtPalavra_AnotadaEN_ = str_replace(" ",
```

```
                        "_", $palavraTermoAnotadoEN);
324
325                     $resultado.=
326                     "
327         <wss_rdfa:$numCVLattes?tags=$ondeLocalCV?tag=
               $txtPalavra_Anotada_>      a                        oa:
               Tag, oa:d4e654, cnt:ContentAsText  ".
               rdf_addTipoAnotacao($tipoEntidadeWiki).";
328                     cnt:chars
                           \"$palavraTermoAnotado\";
329                     cnt:characterEncoding   \"utf-8\";";
330                     $resultado              .=
                           rdf_addPropriedadesSemanticas(
                           $tipoEntidadeWiki,$palavraTermoAnotado)
                           ;
331                     $resultado              .= "\n
                           rdfs:label
                                   \"$palavraTermoAnotado\" ";
332
333                     if(strlen($WikipediaURL)!=0 and
                           $WikipediaURL!="Wikipedia Url"){
334                          $resultado              .=";";
335                          $resultado              .="\n
                                    schema:sameAs   <".
                               rdf_linkWikiPT($WikipediaURL)."
                               >,<".rdf_linkWikiEN(
                               $txtPalavra_AnotadaEN_).">;" ;
336                          $resultado              .="\n
                                    foaf:page
                                    <".rdf_linkWikiPT(
                               $WikipediaURL).">,<".
                               rdf_linkWikiEN(
                               $txtPalavra_AnotadaEN_).">." ;
337
338                          //Se tem wikiPedia:
339                          //monta descri  o dos links
                               semanticos do DBpedia:
340                          $resultado.=
341                          "
342         <".rdf_linkDBpediaPT($txtPalavra_Anotada_).">   a
               oa:SemanticTag, oa:d4e583 ".rdf_addTipoAnotacao(
               $tipoEntidadeWiki).";
```

```
343                         rdfs:label              \"
                               $palavraTermoAnotado\" ;
344                         dbp-prop:label  \"$palavraTermoAnotado\"
                               ;
345                         schema:name     \"$palavraTermoAnotado\"
                               ;";
346                               $resultado              .=
                                  rdf_addPropriedadesSemanticas(
                                  $tipoEntidadeWiki,
                                  $palavraTermoAnotado);
347                               $resultado              .="\n
                                           schema:sameAs   <".
                                  rdf_linkWikiPT($WikipediaURL)."
                                  >,<".rdf_linkWikiEN(
                                  $txtPalavra_AnotadaEN_).">;" ;
348                               $resultado              .="\n
                                           foaf:page
                                           <".rdf_linkWikiPT(
                                  $WikipediaURL).">,<".
                                  rdf_linkWikiEN(
                                  $txtPalavra_AnotadaEN_).">." ;
349
350                               if($txtPalavra_AnotadaEN_!="")
351                               {
352                                     $resultado.="
353       <".rdf_linkDBpediaEN($txtPalavra_AnotadaEN_).">  a
             oa:SemanticTag, oa:d4e583 ".rdf_addTipoAnotacao(
             $tipoEntidadeWiki).";
354                         rdfs:label              \"
                               $palavraTermoAnotado\" ;
355                         dbp-prop:label  \"$palavraTermoAnotado\"
                               ;";
356                               $resultado
                                     .=
                                     rdf_addPropriedadesSemanticas
                                     ($tipoEntidadeWiki,
                                     $palavraTermoAnotado);
357                               $resultado
                                     .="\n           schema:
                                     name     \"
                                     $palavraTermoAnotado\""
                                     ;
```

```
358                                     if(strlen($WikipediaURL)
                                           !=0 and $WikipediaURL!=
                                           "Wikipedia Url"){
359                                           $resultado
                                                        .=" ;";
360                                           $resultado
                                                        .="\n

                                              schema:sameAs
                                                 <".
                                              rdf_linkWikiPT(
                                              $WikipediaURL).
                                              ">,<".
                                              rdf_linkWikiEN(
                                              $txtPalavra_AnotadaEN_
                                              )."> ;" ;
361                                           $resultado
                                                        .="\n
                                                           foaf
                                              :page
                                                        <".
                                              rdf_linkWikiPT(
                                              $WikipediaURL).
                                              ">,<".
                                              rdf_linkWikiEN(
                                              $txtPalavra_AnotadaEN_
                                              )."> ." ;
362                                     }
363
364                                }
365
366                           }else{ //Fim se tem wikiPedia tem DBpedia
367                                $resultado              .=".";
368                           }
369
370
371                           //descri  o semantica do Freebase
372                           if(strlen($FreebaseID)!=0){
373                                $resultado.=
374                                "
375        <http://www.freebase.com/$FreebaseID?lang=pt>   a
                                      oa:SemanticTag, oa:d4e583 ".
```

```
            rdf_addTipoAnotacao($tipoEntidadeWiki).";
376                         rdfs:label              \"
                               $palavraTermoAnotado\" ;
377                         dbp-prop:label  \"$palavraTermoAnotado\"
                               ;
378                         schema:name     \"$palavraTermoAnotado\"
                               ;
379                         schema:sameAs   <".rdf_linkWikiPT(
                               $WikipediaURL).">,<".rdf_linkWikiEN(
                               $WikipediaURL).">;
380                         foaf:page               <".rdf_linkWikiPT
                               ($WikipediaURL).">,<".rdf_linkWikiEN(
                               $WikipediaURL).">.
381                         ";
382                         }
383
384                 }
385
386                 return $resultado."\n\n";
387         }
388
389         function rdf_AnotacoesWss($numCVLattes,$ondeLocalCV,
               $arrayEntidadesAnotadas){
390
391                 $resultado=
392                 "
393         <wss_rdfa:8042937271101537?tags=$ondeLocalCV>";
394
395                 //escrevendo os termos/palavras encontradas no
                       array (resumo, formAcademica, ...):
396                 foreach ($arrayEntidadesAnotadas as $valores)//
                       pegando os valores, entidades, termos do resumo
                       , formacao academ...
397                 {
398                         $tipoEntidadeWiki               = trim(
                               $valores['tipoEntidade']); //pode vir
                               vazio ou "ND" n o definido
399                         $palavraTermoAnotado    = trim($valores['
                               valorEntidade']); //sempre tem
                               conte do
400                         $palavraTermoAnotadoEN  = trim($valores['
                               valorEntidadeEN']); //sempre tem
```

```
                        conte do
401                     $WikipediaURL   = trim($valores['
                        WikipediaURL']); //pode vir vazio
402                     $FreebaseID             = trim($valores['
                        FreebaseID']); //pode vir vazio
403
404                     if($palavraTermoAnotado=="valorEntidade")
                         continue; //cabe alho do array.
405
406                     $resultado.= "  wss:temAnotado  \"
                        $palavraTermoAnotado\" ;
407                                     ";
408             }
409
410             //Tirando a virgula da  ltima  linha e
                adicionando o ponto final.
411             $resultado .=".";
412             $resultado =str_replace(" ;
413                                     .", ".\n\n\n",
                                        $resultado);
414             return $resultado;
415     }
416     function rdf_AnotacoesWssAreaFormAcad($numCVLattes,
            $txtOriginalAreaFormAcad, $ondeLocalCV,
            $arrayEntidadesAnotadas)
417     {
418             $tripla="";
419             $resultadoParcial="";
420             $grauAcademicoConteudo= explode("Grau Acad mico"
                ,$txtOriginalAreaFormAcad);
421
422             $resultadoParcial .="#########Iniciando a Area
                Formacao Academica(formAcad): Mapeamento WSS:\n
                ";
423
424             foreach ($grauAcademicoConteudo as
                $conteudograuAcad)
425             {
426                     if(strlen($conteudograuAcad)<=20)
                        continue;//tem caso que o array[0] tem
                        19 caract.
427                     $nomeGrau = Trim(strstr($conteudograuAcad
```

```
                        , '(', true));
428
429                 foreach ($arrayEntidadesAnotadas as
                        $valores)//pegando os valores,
                        entidades, termos do resumo
430                 {
431                     $tipoEntidadeWiki               =
                            trim($valores['tipoEntidade'])
                        ;
432                     $palavraTermoAnotado    = trim(
                        $valores['valorEntidade']);
433                     $palavraTermoAnotadoEN  = trim(
                        $valores['valorEntidadeEN']);
434                     $WikipediaURL                   =
                            trim($valores['WikipediaURL'])
                        ;
435                     $FreebaseID
                                = trim($valores['
                        FreebaseID']);
436
437                     $txtPalavra_Anotada_ =
                        str_replace(" ", "_",
                        $palavraTermoAnotado);
438                     $txtPalavra_AnotadaEN_ =
                        str_replace(" ", "_",
                        $palavraTermoAnotadoEN);
439
440                     //s o campos XML x RDF Turtle:
441                     $arrayCamposCoReferencia =
                        fun_retornaCamposCoreferenciaXmlRdf
                        ("formAcad");
442                     foreach ($arrayCamposCoReferencia
                            as $linha)
443                     {
444                         //$linha[0]    do XML e
                                $linha[1]    RDF
445                         $textoLattes = $linha[0].
                                "=\"".
                                $palavraTermoAnotado."
                                \"";
446
447                         //existe textoLattes no
```

```
                                        meu textoOriginal
                                        quebrado em
                                        GrauConteudo?
448                                 $encontrouTxtLattes =
                                        strripos(
                                        $conteudograuAcad,
                                        $textoLattes);
449                                 if($encontrouTxtLattes===
                                        false)
450                                         continue;
451                                 else{
452                                         if(strlen(
                                                $WikipediaURL)
                                                !=0 and
                                                $WikipediaURL!=
                                                "Wikipedia Url"
                                                ){
453                                                     //
                                                        resposta
                                                         para o

                                                        arquivo
                                                         turtle
                                                        ...
                                                        triplas
                                                         wss:
454                                                    //campo
                                                        [1]
                                                        a
                                                        propriedade
                                                         RDF
                                                        mapeada
                                                        . ver:
                                                        fun_retornaCampo

455                                             $tripla .= "
                                                        wss:".
                                                $linha[1]."

                                                        <".
                                                rdf_linkDBpediaPT
                                                (
```

```
                                                $txtPalavra_Anotada_
                                                ).">   ;\n";
456                                           }else{
457                                           $tripla .= "
                                                        wss:".
                                                $linha[1]."

                                                        \"
                                                $palavraTermoAnotado
                                                \"        ;\n";
458                                           }
459                                           if(strlen(
                                                $FreebaseID)
                                                !=0){
460                                           $tripla .= "
                                                        wss:".
                                                $linha[1]."

                                                        <http
                                                ://www.freebase
                                                .com/
                                                $FreebaseID?
                                                lang=pt>   ;\n"
                                                ;
461                                           }
462                                   }
463
464                           }//for do array de correferencia
465
466                   }//for do array de palavras anotadas
467
468                   $resultadoParcial .= "<wss_rdfa:
                         $numCVLattes?tags=$ondeLocalCV>
                                 wss:grauAcad           \"
                         $nomeGrau\"   ;\n";
469                   $resultadoParcial .= $tripla;
470
471                   //Tirando a virgula da  ltima  linha e
                         adicionando o ponto final.
472                   $resultadoParcial .=".";
473                   $resultado =str_replace(";\n.", ".\n",
                         $resultadoParcial);
```

```
474
475             }//for nomes dos Graus de Curso
476
477             return $resultado;
478
479     }
480     function rdf_RodapeFim($numCVLattes,
           $anotadoPorIndexSiteWSS,$anotadoPorIndexPessoaSite){
481             $resultado="################ Descri  o Rodap ,
                    dados sistema, extrator e pessoa anotador:
482     <$anotadoPorIndexSiteWSS>       a       foaf:Agent, prov:
           SoftwareAgent;
483                     rdfs:label      \"Lattes Web Sem ntico\"
                            ;
484                     foaf:name       \"Sistema Lattes Web
                           Sem ntico\" .
485     <https://www.textrazor.com/>    a       foaf:Agent, prov:
           SoftwareAgent;
486                     rdfs:label      \"Extrator de Ent.
                           TextRazor\" ;
487                     foaf:name       \"Extrator de Entidade
                           TextRazor\" .
488     <$anotadoPorIndexPessoaSite> a foaf:Person, schema:Person
           , dbp-onto:Person;
489                     rdfs:label      \"Walison Dias da Silva\"
                            ;
490                     foaf:name       \"Walison Dias da Silva
                           \".
491
492     #urn:generated:$numCVLattes.ttl
493     #context:".str_replace("index1.php", "",
           $anotadoPorIndexSiteWSS);
494
495             return $resultado;
496     }
```

APÊNDICE D – Código Fonte:Consulta Sparql:Currículos Cadastrados na base LattesWS

```
1         $TextoExplicativoConsulta=" Quais s o os curr culos que
              est o anotados e cadastrados?";
2
3         $arrayRespConsulta = dtb_executarPesqSparql("PREFIX cv: <
             http://rdfs.org/resume-rdf/cv.rdfs#>
4 SELECT ?nome    ?currLatt
5 WHERE {
6         ?currLatt       a cv:CV ;
7                         cv:aboutPerson  ?nome.
8 }
9 order by ?nome
10 ");
11
12        if($arrayRespConsulta[0]==false){
13                echo $arrayRespConsulta[1];
14                exit;
15        }
16
17
18        $arrayRegistros = dtb_arrayRegistros($arrayRespConsulta
             [1]);//monta um array com a resposta
19        if($arrayRegistros[0]==-1 and $arrayRegistros[1]==-1){//
             linhas e colunas
20                echo "<br>A consulta ($TextoExplicativoConsulta)
                    n o retornou dados.<br>";
21                exit;
22        }
23
24        $numRegistros   =$arrayRegistros[0]."<br>";//quantidade
             de linhas
25        $numColunas             =$arrayRegistros[1]."<br><Br>";//
             quantidade de colunas
26
```

```
27          $NomeColunas     = "Autor Lattes,Curr culo";//passar
               separado por virgula os nomes.
28          $TipoCampoCorpo = "texto,hiperlink";//texto,hiperlink...
29
30          //retorna a tabela em linhas e colunas
31          //tabela possui texto explicativo como um t tulo
32          fun_geraTabelaRespSparql($numRegistros,$numColunas,
               $NomeColunas,$TipoCampoCorpo,$arrayRegistros[2],
               $TextoExplicativoConsulta);
```

APÊNDICE E – Código Fonte:Consulta Sparql:Termo Semântico nos currículos.

```
1   <?php
2
3   $TextoExplicativoConsulta=" Em qual parte dos curr culos
        Lattes existem o termo FUMEC e PUC Rio?";
4
5   $arrayRespConsulta = dtb_executarPesqSparql("PREFIX oa: <
        http://www.w3.org/ns/oa#>
6   PREFIX wss: <http://www.wssistemas.com.br/latteswss/>
7   PREFIX dbp-prop: <http://dbpedia.org/property/>
8   select  ?Curr   ?CurrParte      ?descItem
9   where {
10      {
11              ?CurrParte wss:eParte   ?Curr .
12          ?CurrParte a                oa:Annotation.
13          ?CurrParte oa:hasBody <http://pt.dbpedia.org/
                resource/FUMEC>.
14          <http://pt.dbpedia.org/resource/FUMEC> dbp-prop:
                label ?descItem
15      }UNION{
16              ?CurrParte wss:eParte   ?Curr .
17          ?CurrParte a                oa:Annotation.
18          ?CurrParte oa:hasBody <http://pt.dbpedia.org/
                resource/
                Pontif cia_Universidade_Cat lica_de_Minas_Gerais
                >.
19          <http://pt.dbpedia.org/resource/
                Pontif cia_Universidade_Cat lica_de_Minas_Gerais
                > dbp-prop:label ?descItem.
20      }
21  }");
22
23  if($arrayRespConsulta[0]==false){
24          echo $arrayRespConsulta[1];
25          exit;
26  }
```

```
27
28
29          $arrayRegistros = dtb_arrayRegistros($arrayRespConsulta
               [1]);//monta um array com a resposta
30          if($arrayRegistros[0]==-1 and $arrayRegistros[1]==-1){//
               linhas e colunas
31                echo "<br>A consulta ($TextoExplicativoConsulta)
                     n o retornou dados.<br>";
32                exit;
33          }
34
35          $numRegistros   =$arrayRegistros[0]."<br>";//quantidade
               de linhas
36          $numColunas             =$arrayRegistros[1]."<br><Br>";//
               quantidade de colunas
37
38          $NomeColunas    = "Curr culo Lattes,Parte do Curr culo,
               Termo";//passar separado por virgula os nomes.
39          $TipoCampoCorpo = "hiperlink,texto,texto";//texto,
               hiperlink...
40
41          //retorna a tabela em linhas e colunas
42          //tabela possui texto explicativo como um t tulo
43          fun_geraTabelaRespSparql($numRegistros,$numColunas,
               $NomeColunas,$TipoCampoCorpo,$arrayRegistros[2],
               $TextoExplicativoConsulta);
```

APÊNDICE F – Código Fonte:Consulta Sparql:Termos Semânticos em partes específicas dos currículos Lattes.

```
1  <?php
2
3  require_once("../DataBase.php");
4  require_once("../Geral.php");
5
6  $TextoExplicativoConsulta=" Quais s o os documentos que
       possuem na sua  rea  de Forma  o Acad mica o termo
7                  Engenharia El trica e Banco de Dados e
                       no seu Resumo o termo Tomada de
                       Decis o? Essa consulta
8                  demonstra a efetividade de relacionamento
                        entre as partes dentro de um mesmo
                       curr culo.";
9
10 $arrayRespConsulta = dtb_executarPesqSparql('PREFIX oa: <
       http://www.w3.org/ns/oa#>
11 PREFIX wss: <http://www.wssistemas.com.br/latteswss/>
12 PREFIX dbp-prop: <http://dbpedia.org/property/>
13 select *
14 where  {
15     {
16          ?currPart       wss:pertence    "formAcad".
17                  ?currPart       oa:hasBody              <
                       http://pt.dbpedia.org/resource/
                       Engenharia_El trica>.
18          <http://pt.dbpedia.org/resource/
               Engenharia_El trica>   dbp-prop:label ?desc1
19          }UNION{
20          ?currPart       wss:pertence    "formAcad".
21                  ?currPart       oa:hasBody              <
                       http://pt.dbpedia.org/resource/
                       Banco_de_Dados>.
22          <http://pt.dbpedia.org/resource/Banco_de_Dados>
```

```
                    dbp-prop:label ?desc1
23              }UNION{
24                  ?currPart       wss:pertence    "abstract".
25                          ?currPart       oa:hasBody              <
                               http://pt.dbpedia.org/resource/
                               Tomada_de_Decis o>.
26                  <http://pt.dbpedia.org/resource/
                       Tomada_de_Decis o>     dbp-prop:label ?desc1
27              }
28          }order by ?currPart');
29
30          if($arrayRespConsulta[0]==false){
31                  echo $arrayRespConsulta[1];
32                  exit;
33          }
34
35
36          $arrayRegistros = dtb_arrayRegistros($arrayRespConsulta
               [1]);//monta um array com a resposta
37          if($arrayRegistros[0]==-1 and $arrayRegistros[1]==-1){//
               linhas e colunas
38                  echo "<br>A consulta ($TextoExplicativoConsulta)
                       n o retornou dados.<br>";
39                  exit;
40          }
41
42          $numRegistros   =$arrayRegistros[0]."<br>";//quantidade
               de linhas
43          $numColunas             =$arrayRegistros[1]."<br><Br>";//
               quantidade de colunas
44
45          $NomeColunas    = "Termo,Parte do Curr culo";//passar
               separado por virgula os nomes.
46          $TipoCampoCorpo = "texto,hiperlink";//texto,hiperlink...
47
48          //retorna a tabela em linhas e colunas
49          //tabela possui texto explicativo como um t tulo
50          fun_geraTabelaRespSparql($numRegistros,$numColunas,
               $NomeColunas,$TipoCampoCorpo,$arrayRegistros[2],
               $TextoExplicativoConsulta);
```

APÊNDICE G – Código Fonte:Consulta Sparql:Utilizando o LOD (numberOfPostgraduateStudents) para obter dados.

```
1
2  <?php
3
4  $TextoExplicativoConsulta=" Das universidades que est o na  rea
       de forma  o , qual    a quantidade de
5                  p s -granduandos por universidades? Buscando a
                      quantidade de p s -graduandos no LOD do pt.
                      dbpedia.";
6
7  $arrayRespConsulta = dtb_executarPesqSparql('PREFIX wss: <http://
       www.wssistemas.com.br/latteswss/>
8  PREFIX schema: <http://schema.org/>
9  PREFIX dbp-prop: <http://dbpedia.org/property/>
10 SELECT distinct ?descUniv       ?link   ?numPosGrad
11 WHERE {
12     ?s wss:pertence                          "formAcad".
13         ?s wss:nomeInstituicaoEscola    ?link .
14     ?link a
          schema:University.
15     ?link dbp-prop:label                     ?descUniv
16         FILTER REGEX(str(?link),"pt.dbpedia")
17     Service <http://pt.dbpedia.org/sparql>{
18         ?link   <http://dbpedia.org/ontology/
              numberOfPostgraduateStudents>      ?numPosGrad
19     }
20 }
21 order By ?descUniv ?qntdade');
22
23 if($arrayRespConsulta[0]==false){
24         echo $arrayRespConsulta[1];
25         exit;
```

```
26 }
27
28
29 $arrayRegistros = dtb_arrayRegistros($arrayRespConsulta[1]);//
      monta um array com a resposta
30 if($arrayRegistros[0]==-1 and $arrayRegistros[1]==-1){//linhas e
      colunas
31         echo "<br>A consulta ($TextoExplicativoConsulta) n o
              retornou dados.<br>";
32         exit;
33 }
34
35 $numRegistros   =$arrayRegistros[0]."<br>";//quantidade de linhas
36 $numColunas             =$arrayRegistros[1]."<br><Br>";//
      quantidade de colunas
37
38 $NomeColunas    = "Universidade,Link Dbpedia, Quantidade P s -
      Graduandos";//passar separado por virgula os nomes.
39 $TipoCampoCorpo = "texto,hiperlink,texto";//texto,hiperlink...
40
41 //retorna a tabela em linhas e colunas
42 //tabela possui texto explicativo como um t tulo
43 fun_geraTabelaRespSparql($numRegistros,$numColunas,$NomeColunas,
      $TipoCampoCorpo,$arrayRegistros[2],$TextoExplicativoConsulta);
```

APÊNDICE H – Código Fonte:Consulta Sparql:Utilizando o LOD(populacaoEstimada,clima e tipoGov).

```
1   <?php
2
3   $TextoExplicativoConsulta=" Quais s o as anota  es
        encontradas nos curr culos lattes que s o
        classificados
4                   como um pa s e desses, represente a sua
                        quantidade populacional, o clima e seu
                        tipo de governo.
5                   Buscando as informa  es quantidade
                        populacional, o clima e seu tipo de
                        governo no LOD do pt.dbpedia.";
6
7   $arrayRespConsulta = dtb_executarPesqSparql('PREFIX
        schema: <http://schema.org/>
8   PREFIX oa: <http://www.w3.org/ns/oa#>
9   PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
10  PREFIX dbp-prop: <http://pt.dbpedia.org/property/>
11  select *
12  where {
13            ?s     a       schema:Country.
14            ?s     a       oa:SemanticTag.
15            Service <http://pt.dbpedia.org/sparql>{
16                      ?s rdfs:label ?Pais   .
17                      ?s dbp-prop:popula  oEstimada       ?
                            popTotal       .
18                      ?s dbp-prop:clima
                                    ?clima          .
19                      ?s dbp-prop:tipoGoverno
                                    ?tipoGov        .
20            }
21  }');
22
23  if($arrayRespConsulta[0]==false){
```

```
24                  echo $arrayRespConsulta[1];
25                  exit;
26          }
27
28
29          $arrayRegistros = dtb_arrayRegistros($arrayRespConsulta
               [1]);//monta um array com a resposta
30          if($arrayRegistros[0]==-1 and $arrayRegistros[1]==-1){//
               linhas e colunas
31                  echo "<br>A consulta ($TextoExplicativoConsulta)
                       n o retornou dados.<br>";
32                  exit;
33          }
34
35          $numRegistros   =$arrayRegistros[0]."<br>";//quantidade
               de linhas
36          $numColunas             =$arrayRegistros[1]."<br><Br>";//
               quantidade de colunas
37
38          $NomeColunas    = "Pa s Dbpedia,Nome,Quantidade
               Populacional,Clima,Tipo de Governo";//passar separado
               por virgula os nomes.
39          $TipoCampoCorpo = "hiperlink,texto,texto,texto,texto";//
               texto,hiperlink...
40
41          //retorna a tabela em linhas e colunas
42          //tabela possui texto explicativo como um t tulo
43          fun_geraTabelaRespSparql($numRegistros,$numColunas,
               $NomeColunas,$TipoCampoCorpo,$arrayRegistros[2],
               $TextoExplicativoConsulta);
```